Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

Doktor-Ingenieur
(Dr.-Ing.)

im Fachgebiet

Informatik

Vorgelegt
von Herr Dipl.-Ing. (FH) Andreas Niekler
geboren am 21. Juli 1979 in Torgau.

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Gerhard Heyer (Universität Leipzig, Deutschland)
2. Prof. Dr. Christiane Fellbaum (Princeton University, USA)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 13. Januar 2016 mit dem Gesamtprädikat *magna cum laude*.

Vorwort

Die Idee für die Dissertation entstand am Lehrgebiet Elektronische Mediensystemtechnik der Fakultät Medien an der HTWK Leipzig und wurde gemeinsam mit Prof. Dr. Uwe Kulisch, der dieses Lehrgebiet betreut, entwickelt. Im Rahmen eines kooperativen Promotionsverfahrens wurde die Arbeit an der Dissertation im Jahr 2009 aufgenommen. Die Betreuung seitens einer Universität wurde durch Prof. Dr. Gerhard Heyer vom Lehrstuhl für Automatische Sprachverarbeitung der Universität Leipzig übernommen. Während der Betreuung durch beide Lehrstühle wurden unterschiedliche Anwendungsszenarien des Forschungsgegenstandes deutlich. Seitens der HTWK liegt der Fokus auf kommunikationswissenschaftlichen und journalistischen Anwendungen. Der Schwerpunkt am Lehrstuhl der Universität Leipzig liegt im Gebiet der Automatischen Sprachverarbeitung mit Anwendungen in den Digital Humanities. Beide Sichtweisen wurden hinsichtlich der damit verbundenen Anforderungen eingebunden, sodass der kooperative Charakter in der Arbeit wiederzufinden ist. Die Entscheidung in deutscher Sprache zu schreiben, ist durch diese Kooperation beeinflusst, da die Betreuung an zwei deutschen Hochschulen übernommen wurde. Für die Betreuung der Arbeit möchte ich Prof. Dr. Uwe Kulisch und Prof. Dr. Gerhard Heyer herzlich danken.

Die Umsetzung des Dissertationsprojekts wurde seitens der HTWK und der Univerität Leipzig sehr unterstützt. Ein Förderprojekt für die Disseration, herausgegeben durch die Sächsische Aufbaubank (SAB), wurde durch das Forschungs- und Transferzentrum Leipzig e.V. (FTZ) initiiert. Mein besonderer Dank gilt den dortigen Mitarbeitern und besonders Peggy Stöckigt und Olga Lüders, die den Antrag, die Finanzierung und die Projektdurchführung großartig unterstützt haben. Weiterhin möchte ich Gregor Wiedemann danken, der das Digital Humanties Projekt, in dessen Rahmen ich die Arbeit an der Dissertation beenden konnte, initiiert, beantragt und wesentlich gestaltet hat.

Zusammenfassung

Im Bereich der medienwissenschaftlichen Inhaltsanalyse stellt die Themenanalyse einen wichtigen Bestandteil dar. Für die Analyse großer digitaler Textbestände hinsichtlich thematischer Strukturen ist es deshalb wichtig, das Potential automatisierter computergestützter Methoden zu untersuchen. Dabei müssen die methodischen und analytischen Anforderungen der Inhaltsanalyse beachtet und abgebildet werden, welche auch für die Themenanalyse gelten. In dieser Arbeit werden die Möglichkeiten der Automatisierung der Themenanalyse und deren Anwendungsperspektiven untersucht. Dabei wird auf theoretische und methodische Grundlagen der Inhaltsanalyse und auf linguistische Theorien zu Themenstrukturen zurückgegriffen, um Anforderungen an eine automatische Analyse abzuleiten. Den wesentlichen Beitrag stellt die Untersuchung der Potentiale und Werkzeuge aus den Bereichen des Data- und Text-Mining dar, die für die inhaltsanalytische Arbeit in Textdatenbanken hilfreich und gewinnbringend eingesetzt werden können. Weiterhin wird eine exemplarische Analyse durchgeführt, um die Anwendbarkeit automatischer Methoden für Themenanalysen zu zeigen. Die Arbeit demonstriert auch Möglichkeiten der Nutzung interaktiver Oberflächen, formuliert die Idee und Umsetzung einer geeigneten Software und zeigt die Anwendung eines möglichen Arbeitsablaufs für die Themenanalyse auf. Die Darstellung der Potentiale automatisierter Themenuntersuchungen in großen digitalen Textkollektionen in dieser Arbeit leistet einen Beitrag zur Erforschung der automatisierten Inhaltsanalyse.

Ausgehend von den Anforderungen, die an eine Themenanalyse gestellt werden, zeigt diese Arbeit, mit welchen Methoden und Automatismen des Text-Mining diesen Anforderungen nahe gekommen werden kann. Zusammenfassend sind zwei Anforderungen herauszuheben, deren jeweilige Erfüllung die andere beeinflusst. Zum einen ist eine schnelle thematische Erfassung der Themen in einer komplexen Dokumentsammlung gefordert, um deren inhaltliche Struktur abzubilden und um Themen kontrastieren zu können. Zum anderen müssen die Themen in einem ausreichenden Detailgrad abbildbar sein, sodass eine Analyse des Sinns und der Bedeutung der Themeninhalte möglich ist. Beide Ansätze haben eine methodische Verankerung in den quantitativen und qualitativen Ansätzen der Inhaltsanalyse. Die Arbeit diskutiert diese Parallelen und setzt automatische Verfahren und Algorithmen mit den Anforderungen in Beziehung. Es können Methoden aufgezeigt werden, die eine semantische und damit thematische Trennung der Daten erlauben und einen abstrahierten Überblick über große Dokumentmengen schaffen. Dies sind Verfahren wie Topic-Modelle oder clusternde Verfahren. Mit Hilfe dieser Algorithmen ist es möglich, thematisch kohärente Untermengen in Dokumentkollektion zu erzeugen und deren thematischen Gehalt für Zusammenfassungen bereitzustellen. Es wird gezeigt, dass die Themen

trotz der distanzierten Betrachtung unterscheidbar sind und deren Häufigkeiten und Verteilungen in einer Textkollektion diachron dargestellt werden können. Diese Aufbereitung der Daten erlaubt die Analyse von thematischen Trends oder die Selektion bestimmter thematischer Aspekte aus einer Fülle von Dokumenten. Diachrone Betrachtungen thematisch kohärenter Dokumentmengen werden dadurch möglich und die temporären Häufigkeiten von Themen können analysiert werden. Für die detaillierte Interpretation und Zusammenfassung von Themen müssen weitere Darstellungen und Informationen aus den Inhalten zu den Themen erstellt werden. Es kann gezeigt werden, dass Bedeutungen, Aussagen und Kontexte über eine Kookurrenzanalyse im Themenkontext stehender Dokumente sichtbar gemacht werden können. In einer Anwendungsform, welche die Leserichtung und Wortarten beachtet, können häufig auftretende Wortfolgen oder Aussagen innerhalb einer Thematisierung statistisch erfasst werden. Die so generierten Phrasen können zur Definition von Kategorien eingesetzt werden oder mit anderen Themen, Publikationen oder theoretischen Annahmen kontrastiert werden. Zudem sind diachrone Analysen einzelner Wörter, von Wortgruppen oder von Eigennamen in einem Thema geeignet, um Themenphasen, Schlüsselbegriffe oder Nachrichtenfaktoren zu identifizieren. Die so gewonnenen Informationen können mit einem „close-reading“ thematisch relevanter Dokumente ergänzt werden, was durch die thematische Trennung der Dokumentmengen möglich ist. Über diese methodischen Perspektiven hinaus lassen sich die automatisierten Analysen als empirische Messinstrumente im Kontext weiterer hier nicht besprochener kommunikationswissenschaftlicher Theorien einsetzen. Des Weiteren zeigt die Arbeit, dass grafische Oberflächen und Software-Frameworks für die Bearbeitung von automatisierten Themenanalysen realisierbar und praktikabel einsetzbar sind. Insofern zeigen die Ausführungen, wie die besprochenen Lösungen und Ansätze in die Praxis überführt werden können.

Wesentliche Beiträge liefert die Arbeit für die Erforschung der automatisierten Inhaltsanalyse. Die Arbeit dokumentiert vor allem die wissenschaftliche Auseinandersetzung mit automatisierten Themenanalysen. Während der Arbeit an diesem Thema wurden vom Autor geeignete Vorgehensweisen entwickelt, wie Verfahren des Text-Mining in der Praxis für Inhaltsanalysen einzusetzen sind. Unter anderem wurden Beiträge zur Visualisierung und einfachen Benutzung unterschiedlicher Verfahren geleistet. Verfahren aus dem Bereich des Topic Modelling, des Clustering und der Kookkurrenzanalyse mussten angepasst werden, sodass deren Anwendung in inhaltsanalytischen Anwendungen möglich ist. Weitere Beiträge entstanden im Rahmen der methodologischen Einordnung der computergestützten Themenanalyse und in der

Definition innovativer Anwendungen in diesem Bereich. Die für die vorliegende Arbeit durchgeführte Experimente und Untersuchungen wurden komplett in einer eigens entwickelten Software durchgeführt, die auch in anderen Projekten erfolgreich eingesetzt wird. Um dieses System herum wurden Verarbeitungsketten, Datenhaltung, Visualisierung, grafische Oberflächen, Möglichkeiten der Dateninteraktion, maschinelle Lernverfahren und Komponenten für das Dokumentretrieval implementiert. Dadurch werden die komplexen Methoden und Verfahren für die automatische Themenanalyse einfach anwendbar und sind für künftige Projekte und Analysen benutzerfreundlich verfügbar. Sozialwissenschaftler, Politikwissenschaftler oder Kommunikationswissenschaftler können mit der Softwareumgebung arbeiten und Inhaltsanalysen durchführen, ohne die Details der Automatisierung und der Computerunterstützung durchdringen zu müssen.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die Benutzung digitaler Textarchive und -quellen erlaubt Zugriff auf darin manifestiertes Wissen und Erkenntnis. In zunehmendem Maße werden Archive, Bibliotheken und redaktionelle Inhalte digital verfügbar. Die maschinelle Verarbeitung von digitalem Text mithilfe der Informatik erleichtert und optimiert die Arbeit mit digitalen Textarchiven. Allerdings ist die analytische Arbeit mit digitalen Quellen ungleich komplizierter als die digitale Archivierung und Suche. Zum einen sind Volltexte nicht immer verfügbar und zum anderen ist das enthaltene Wissen meist nicht vorstrukturiert. Die Inhalte sind nicht effizient verfügbar. Die Analyse von Texten ist aber wichtiges und grundlegendes Arbeitsmittel vieler Tätigkeiten im Bereich des Journalismus und der Wissenschaft, insbesondere der Soziologie und der Kommunikationswissenschaft, weil viele Fachrichtungen bei der Arbeit auf Inhaltsanalysen setzen, um mit dem in Texten und anderen Medien enthaltenen Wissen, Antworten auf ihre Fragestellungen zu finden. Der Inhaltsanalyse geht es um die Erhebung empirischer Daten mittels einer strukturierten und offengelegten Suchstrategie, die aus materialisierter Kommunikation gewonnen werden können (Früh, 2007, vgl. S. 147). In dieser Arbeit wird untersucht, welche Potentiale und Werkzeuge aus den Bereichen des Data- und Text-Mining für die inhaltsanalytische Arbeit in Textdatenbanken hilfreich und gewinnbringend eingesetzt werden können. Dabei konzentriert sich die Arbeit auf die Themenanalyse, einen Teilbereich der Inhaltsanalyse, welcher in Kapitel 2 konkretisiert und zur Definition grundlegender Anforderungen führt. Die Darstellung der Potentiale automatisierter Themenuntersuchungen in großen digitalen Textkollektionen in dieser Arbeit leistet dabei einen Beitrag zur Erforschung der automatisierten Inhaltsanalyse. Dabei setzt die Inhaltsanalyse einerseits in größeren repräsentativen Mengen von Inhalten auf die quantitative Analyse manifester Messgrößen, wie beispielsweise eines bestimmten Vokabulars. Der Gesamtzusammenhang der Inhalte

steht im Hintergrund, da lediglich einzelne Messgrößen bewertet werden. Andererseits kann ein qualitativer Zugang in wenigen Dokumenten erfolgen, um vorliegende Inhalte durch einen detaillierten Zugriff auf alle inhaltlichen Zusammenhänge zu verstehen. Bei der analytischen Arbeit in Textarchiven gibt es verschiedene Dimensionen, die für die Analysten eine Rolle spielen. So gliedern sich Textsammlungen in verschiedene Dokumente auf, die unterschiedliche thematische Bezüge haben und zu verschiedenen Zeiten oder an unterschiedlichen Orten veröffentlicht werden. Beispielsweise kann die Veränderung von Inhalten gezeigt werden, wenn Inhalte veröffentlichter Dokumente unterschiedlicher Zeiträume analysiert werden. Vorstellbar ist die Beobachtung der lokalen Unterschiede, wenn die Inhalte zwischen verschiedenen Regionen oder Ländern verglichen werden sollen. Diese inhaltsanalytische Aufgabe steht im Gegensatz zu reinen Retrieval-Aufgaben, also der gezielten Suche nach definierten Inhalten. Durch eine umfassende wissenschaftliche Diskussion und die damit verbundene Entwicklung der Inhaltsanalyse ist eine mächtige Analysemethode zur Bearbeitung von Fragestellungen entstanden. Anwendungen der Inhaltsanalyse lassen sich in der Kommunikations und Medienwissenschaft, im Journalismus, den Geschichtswissenschaften, der Betriebswirtschaft, den Politikwissenschaften, dem Marketing u. v. a. aufzeigen. Die Fragestellungen reichen dabei von Themenanalysen oder Argumentationsanalysen und dem Wissensmanagement bis zur Auswertung der Kundenzufriedenheit eines Unternehmens.

Eine besondere Rolle für verschiedene Fragestellungen spielen dabei Inhalte, die aus einem redaktionellem Kontext stammen, wie Nachrichtenartikel aus Tageszeitungen und Nachrichtenmagazinen. Diese Inhalte referenzieren auf eine angenommene Realität, die der medialen Öffentlichkeit journalistisch interpretiert und kommentiert wird. Solche Quellen enthalten demnach Ereignisse, Diskurse, Themen oder Personen, die zu gegebenen Zeiten in einer sozialen Öffentlichkeit stattfinden. Diese Inhalte können durch eine Inhaltsanalyse zugänglich gemacht werden und beispielsweise für die Beurteilung gesellschaftlicher Prozesse herangezogen werden. So ist es besonders diese Textsorte, welche manifestiertes Wissen über längere Zeit und tagesaktuell sichtbar macht. Aus diesem Grund spielt die Untersuchung geeigneter Analysemethoden für diese Textsorte eine zentrale Rolle in der vorliegenden Arbeit.

Bei der Analyse von Nachrichten, Kommentaren oder anderen Meldungen können verschiedene Herangehensweisen eine Rolle spielen, die zu unterschiedlichen Lösungen und Problemen führen. Grundsätzlich können Inhaltsanalysen textuell vorliegender Nachrichten retrospektiv oder prospektiv untersucht werden. Bei einer retrospektiven Analyse werden alle zur Verfügung stehenden Materialien untersucht, die für die Ge-

nerierung oder die Überprüfung von Fragestellungen von Belang sind. Das bedeutet insbesondere, dass in Nachrichtenarchiven die Daten selektiert werden, die bereits vorliegen und geeignet für die Überprüfung der Fragestellung sind. Im Gegensatz dazu sind prospektive Studien begleitend und das auszuwertende Material muss erst erhoben werden. Beispielsweise wäre die begleitende Analyse der Themenstruktur von Tageszeitungen eine prospektive Analyse, da ständig neue Daten erhoben werden müssen. In diesem Zusammenhang müssen bereits vorhandene Daten aus bestehenden Textkollektionen selektiert werden und neue Daten ständig in die Kollektion eingebracht und für die weitere Verarbeitung und Selektion indiziert werden.

Da Kollektionen mit Nachrichtenartikeln sehr große Mengen an Texten enthalten, muss bei der Selektion auf technische Hilfsmittel zurückgegriffen werden, die innerhalb der Textdaten eine sinnvolle Selektion oder Erweiterung erlauben. Als weit verbreitetes Hilfsmittel wird hier die Volltextsuche und Indizierung genutzt, um einen Zugang zu ermöglichen, der über den Inhalt der Texte funktioniert. Weiterhin können, falls vorhanden, Metadaten einzelner Dokumente genutzt werden, um Selektionen über Ressorts, Personen oder Publikationen zu ermöglichen. Die valide Selektion relevanter Inhalte und eine nachfolgende Analyse ist mit reinen Volltextdaten durchaus möglich, aber mit Problemen behaftet. Durch die reine Selektion anhand von definierten Schlüsselwörtern und der damit verbundenen Ambiguitäten kann nicht sicher entschieden werden, ob eine so gefundene Menge von Dokumenten wirklich zur Fragestellung passt oder andere Zusammenhänge enthält. Der Volltextindex bildet als reine Such- und Indizierungsfunktion nicht ab, wie sich beispielsweise Thematisierungen inhaltlich ändern. Werden innerhalb einer Thematisierung Akteure oder Referenzen auf Ereignisse verändert, so läuft eine reine Volltextsuche Gefahr, diese Änderungen bei der Selektion der Untersuchungsmenge oder Grundgesamtheit nicht zu berücksichtigen und unvollständig zu sein. Aus diesem Grund muss die analytische Arbeit in großen Textkollektionen weitere Methoden berücksichtigen, die

- inhaltliche Strukturen analysieren,
- inhaltsanalytische Fragestellungen abbilden können,
- zeitliche Zusammenhänge bilden können,
- referenzierte Orte, Personen oder Institutionen integrieren und
- mit der Textmenge umgehen können.

Sollen unbekannte bzw. neue Phänomene oder Zusammenhänge, bei denen noch keine Theoriebildung erfolgt ist, untersucht werden, so ist oft nicht klar, wie diese in den

Dokumenten identifiziert werden können. Durch diese Unklarheit ist es nicht möglich, geeignete Texte zu selektieren. Die Auswahl von Schlüsselwörtern fällt schwer, da die Bildung von Hypothesen und einer Theorie noch nicht abgeschlossen ist. Aus diesem Grund muss es möglich sein, die Inhalte nach verschiedenen Kriterien zu explorieren. Strukturen müssen sichtbar gemacht werden, die nicht im Voraus definiert sind. Dies dient der Theoriebildung und die Analysten können sich mit dem Datenbestand und dessen Inhalt vertraut machen. Innerhalb von Nachrichtenmeldungen sind dabei Themenstrukturen von Interesse, da diese das öffentliche Interesse und das öffentliche Geschehen abbilden. Im Gegensatz dazu werden Diskurse und Argumentationen in den textuellen Meldungen und Kommentaren mitgeführt, die oft für Inhaltsanalysen die eigentliche Aussage über ein bestimmtes Phänomen liefern. So führt Früh aus, dass die Analyse von Diskursen und Argumenten hohe Anforderungen an eine Inhaltsanalyse stellt und mehrere Indikatoren innerhalb der Texte untersucht werden müssen (Früh, 2007, vgl. S. 85). Durch diese Komplexität ist es umso wichtiger, dass die Auswahl der zu analysierenden Dokumente aus einer Textkollektion sehr genau und vollständig passiert, da diese Vorarbeit eine wichtige Voraussetzung für die Qualität der Analyse ist. Deshalb ist es wichtig, Möglichkeiten der Exploration zu haben, um Relevantes von nicht Relevantem zu trennen. Dies kann beispielsweise die thematische Einschränkung einer Dokumentmenge sein. Aus diesem Grund müssen für hypothesengeleitete Analysen in großen Textkollektionen Methoden zur Verfügung stehen, die Textstrukturen oder Themenstrukturen erkennen und anhand derer weitere Analysen und Dokumentmengen vorbereitet werden können.

Der analytische Umgang mit Daten wird unter dem Begriff Data-Mining zusammengefasst. Data-Mining ist eine Disziplin, bei der große Datenmengen zur Wissensgewinnung durch Algorithmen analysiert werden. Dabei konzentriert sich Data-Mining auf numerische Datenbankdaten, wie beispielsweise den Verkaufszahlen in einer Region. Im Data-Mining werden solche Rohdaten genutzt, um Muster, Trends oder Abweichungen zu messen. Die Methoden des Data-Mining generieren demnach Ergebnisse, die angewandt auf die analytische Arbeit in Textarchiven eine Arbeitsoptimierung versprechen. Die Übertragung der Data-Mining-Methoden auf Textdatenbanken wird unter dem Begriff Text-Mining zusammengefasst (Heyer, 2006, vgl. S. 4). Unter Zunahme der Methoden der Verarbeitung natürlicher Sprache aus der Informatik ergibt sich hier ein wissenschaftliches Untersuchungsfeld, dass bereits enormes Potential entwickelt hat. Dennoch stützt sich die inhaltliche Arbeit und die Forschung in Textarchiven, seien sie digital oder nicht, oft auf den manuellen Umgang mit den Dokumenten. Dabei wird versucht, einen repräsentativen Querschnitt

der Grundgesamtheit aller Dokumente in einem Archiv manuell zu untersuchen. Die Unterstützung durch die elektronische Datenverarbeitung beschränkt sich oft auf Such- und Verwaltungsfunktionen der Dokumente. Die Analyse selbst wird mit einer konkreten Operationalisierung meist manuell durchgeführt. Oft ist es für die zu untersuchenden Textmengen aus ökonomischen Gründen nicht möglich, mehrere Operationalisierungen oder Untersuchungsmethoden zu erproben. Dennoch gibt es hierfür in der Informatik Ansätze, die vielversprechende Ergebnisse liefern, um die digitalen Inhalte für Untersuchungen in verschiedenen Disziplinen zugänglich zu machen und die manuelle Arbeit mit Dokumenten zu entlasten.[1]

1.1 Ausgangslage

Die Analyse von Textdatenbanken kann unter zwei verschiedenen Bedingungen erfolgen. Sind neue Themen oder Änderungen an einer Textdatenbank bei einer prospektiven Analyse signifikant, so sollen sie gemessen und beurteilt werden können. Sowohl bei prospektiven und retrospektiven Untersuchungen kann bereits bestehender Inhalt mit einem zeitlichen Bezug betrachtet und analysiert werden, um die Dynamik des in den Texten enthaltenen Wissens zu untersuchen. Untersuchungen die einen Zeitpunkt mit einem anderen vergleichen, werden diachron genannt und stehen im Gegensatz zu Untersuchungen, die nicht auf zeitliche Bezüge beschränkt sind. In zwei Beispielen soll die praktische Relevanz dieser Unterscheidungen für unterschiedliche Disziplinen erläutert werden.

In der Kommunikations- und Medienwissenschaft, genauer in der empirischen Kommunikationsforschung, stellt die Beurteilung der Berichterstattung über ein Thema eine besondere Aufgabe dar. In verschiedenen Quellen werden Belegstellen für ein Thema gesucht, um zu messen, wie sich das Thema in einer bestimmten zeitlichen Abfolge in dem untersuchten Medium verhält. Dabei stellt die zeitliche Einordnung einen wesentlichen Bestandteil der Untersuchung dar, da mitunter herausgefunden werden soll, welche Ereignisse und Entwicklungen Einfluss auf die Berichterstattung oder die Themenaufmerksamkeit hatten. Inhaltsanalysen mit einem Fokus auf Thematisierungen in textuellen Nachrichtenmedien sind demnach diachrone Textanalysen. So kann beispielsweise die nachhaltige Wirkung von Ereignissen auf Themen durch retrospektive Analysen in Nachrichtentexten gemessen werden. Auf der anderen Seite

[1] In Scharkow (2012) werden Methoden des Data-Mining und des maschinellen Lernens für die Anwendung in der Inhaltsanalyse untersucht. Weiterhin zeigen die Aufsatzsammlungen in West (2001) und Sommer (2014) das Potential von computergestützten Ansätzen.

können mit prospektiven Analysen Trends für Themen abgeleitet werden, ohne die Ereignisse oder Einflussfaktoren zu analysieren, die dafür verantwortlich sind.

Ein weiteres Beispiel stellt die Analyseanforderungen im journalistischen Umfeld heraus. Eine Redaktion befindet sich immer im Spannungsfeld zwischen aktuellen Informationen, eigenen Inhalten, der öffentlichen Meinung und der ständigen Beobachtung der Medienrezipienten. Durch die mittlerweile etablierte Beteiligung der Rezipienten am Inhalt, sei es in Form von Rückkanälen bzw. Interaktivität oder von eigenen Beiträgen, können diese direkt und unmittelbar beobachtet werden. Um das Umfeld, die Wirkungen und den Erfolg einer Redaktion zu verstehen, müssen ständige Medienresonanzanalysen durchführt werden. Redaktionen arbeiten mit rechnergestützten Systemen, weshalb heute ein Großteil der Inhalte digital vorliegt. Wichtig für die Arbeit der Redaktionen ist in diesem Szenario, dass Trends in der Berichterstattung oder aufkommende Themen rechtzeitig aufgegriffen werden können und ein Überblick über die Inhalte bewahrt werden kann. Das widerspiegeln der sogenannten Medienagenda und der Erwartungen der Rezipienten ist in diesem Fall ein ständiger Prozess, der darauf angewiesen ist, aktuelle Entwicklungen in den Textdatenbanken, in diesem Fall Nachrichtentexten, hervorzuheben und sichtbar zu machen.

Zusammenfassend kann gesagt werden, dass die Themenanalyse einen wichtigen Bestandteil fast aller Inhaltsanalysen darstellt, um die Themen selbst eingehender zu untersuchen oder um die Textanalysen an einem bestimmten Thema zu orientieren. Dabei spielt es eine Rolle, ob die analysierte Textmenge ständig wächst, also veränderlich ist, und prospektiv oder als abgeschlossene Menge retrospektiv untersucht wird. Zusätzlich können bei geeigneten Textquellen zeitliche Unterschiede untersucht werden. Existiert in einer Textsammlung keine diachrone Unterscheidung, beispielsweise über Zeitstempel einzelner Dokumente, so ist die Textmenge nur als Ganzes zu untersuchen und die Untersuchung von temporären Trends ist nicht möglich.

1.2 Problemstellung und Ziele

Die vorliegende Arbeit untersucht, wie Verfahren des Text-Mining die inhaltlich-thematische Auswertung in Textarchiven unterstützt und welchen Anforderungen diese Analyse genügen muss. Dabei konzentrieren sich die Untersuchungen auf publizierte Inhalte von Redaktionen (redaktioneller Inhalt), wie Nachrichtenartikel aus Tageszeitungen, Journalen und Magazinen. In den Inhalten der Quellen müssen Themen erkannt und über längere Zeiträume beobachtet werden. Die diachrone Analyse und die nicht-diachrone Analyse werden im Zusammenhang mit verschiedenen Textsorten und maschinellen Lernverfahren untersucht. Dabei ist die Eignung der Verfahren für

prospektive und retrospektive Analysen über eine geeignete Repräsentation und Verarbeitung der Daten herstellbar. Es sollen Verfahren verwendet und evaluiert werden, welche die Relevanz, Aktivität, Kontexte und Veränderungen von Themen über die Zeit verfolgen. Besonders aufschlussreich sind dabei die Vorgänge und Merkmale, die das öffentliche Interesse eines Themas begleiten. In einer exemplarischen Untersuchung wird deshalb untersucht, wie diese Veränderungen am sprachlichen Kontext und der Häufigkeit eines Themas beobachtet werden können.

Die Aufnahme bestimmter Inhalte und die Relevanz von Artikeln spiegelt die Medienöffentlichkeit wider. An der Analyse und an Rückschlüssen auf reale Phänomene anhand der Medienöffentlichkeit sind viele Disziplinen interessiert. Ob ein Thema inhaltlich interessant ist, wird von verschiedenen Faktoren bestimmt. Dazu gehören die sogenannten Nachrichtenfaktoren und die Verfügbarkeit der Informationen. Die Daten, welche mit in dieser Arbeit untersuchten Verfahren erzeugt werden, sollen solche Faktoren bestimmbar und bewertbar machen. Aus der Gesamtlage der so entstehenden Daten können Redaktionen oder Forschergruppen Aussagen und Interpretation über Thematisierungen erstellen. Dies erhöht die Effektivität und die Qualität einer Analyse von Textdaten.

Die Möglichkeit, Kontextveränderungen zu bestimmen, die zu einem erhöhten Informationsinteresse der Öffentlichkeit führen, muss bei einer Untersuchung der Thematisierungen vorhanden sein. Demnach ergeben sich für die Arbeit Anforderungen wie

- die Identifikation von Verfahren, mit denen themenbasierte Inhaltsanalysen möglich sind,
- die Evaluation und Beurteilung dieser Verfahren,
- die Identifikation von Merkmalen eines Themas, die durch die Verfahren messbar werden wie
  - die Beobachtung über die Zeit,
  - die Beobachtung des Themenanteils am Textarchiv und
  - die Beobachtung von weiteren Informationen in einem Thema, wie beispielsweise Nennungen von Personen und
- die Identifikation von Wortkontexten innerhalb eines Themas, die zur Beurteilung desselben Themas beitragen.

1.3 Aufbau der Arbeit

In einem Grundlagenteil, der sich in Kapitel 2 anschließt, wird zunächst erläutert, welche Anforderungen beachtet werden müssen, um methodisch korrekte Inhaltsanalysen durchzuführen. Diese Einführung dient der Vorstellung der Methoden der Inhaltsanalyse, welche in den Kommunikations- und Sozialwissenschaften etabliert sind. Diese Einführung soll eine Schnittstelle zur kommunikationswissenschaftlichen Anwendung der Inhaltsanalyse herstellen. Weiterhin werden in diesem Kapitel Methoden für die computergestützte Verarbeitung von Textdokumenten vorgestellt, die eine Grundlage für die automatische Durchführung von Inhaltsanalysen darstellen. Aus den Ausführungen in Kapitel 2 ergeben sich damit Anforderungen und Fragestellungen für Umsetzung automatisierter Inhalts- bzw. Themenanalysen. In Abschnitt 2.4 werden die Forschungsfragen und Anforderungen der Arbeit noch einmal hinsichtlich der Grundlagen konkretisiert.

Kapitel 3 spricht verschiedene Verfahren an, die sich für die Themenanalyse in elektronischen Dokumenten eignen. Die Arbeit konzentriert sich dabei auf die Verwendung von Verfahren, die ohne Training und Vorwissen arbeiten können. Die manuelle Erstellung von Trainingsdaten ist aufwändig. Da die Inhalte der nachrichtenorientierten Quellen sehr veränderlich sind, ist eine Anpassungen der Trainingsdaten in kurzen Abständen nötig.

In Kapitel 4 werden die untersuchten Verfahren in exemplarischen Analysen implementiert, erprobt und evaluiert. Dabei werden sowohl die Analyse inhaltlicher Aspekte, als auch die Auswertung von Zeitreihen durchgeführt. Die Evaluierung der Verfahren spielt eine wichtige Rolle, um deren Eignung und Anwendbarkeit für Inhaltsanalysen zu zeigen. In Kapitel 5 schließt sich eine Diskussion der Ergebnisse und Erkenntnisse an. Dabei werden konkrete Anforderungen und Forschungsfragen, die hinsichtlich der methodischen Grundlagen der Inhaltsanalyse in Abschnitt 2.4 aufgestellt werden, diskutiert und beantwortet.

1.4 Wesentliche Beiträge

Die hier vorliegende Arbeit dokumentiert vor allem die wissenschaftliche Auseinandersetzung mit automatisierten Themenanalysen. Während der Arbeit an diesem Thema mussten unterschiedliche Softwarekomponenten erstellt und Verfahren angepasst werden. Zudem werden methodologische Vorgehensweisen definiert. Diese Beiträge des Autors werden in aktuellen und künftigen wissenschaftlichen Projekten gewinnbringend eingesetzt (Wiedemann u. a., 2013). Die Ingenieurs- und Designleis-

tungen werden im Rahmen der wissenschaftlichen Auseinandersetzung im weiteren Text aber aus Gründen der Stringenz in den Hintergrund treten. Aus diesem Grund sollen die wissenschaftlichen und technischen Beiträge des Autors an dieser Stelle vorab zusammengefasst und vorgestellt werden.

In dieser Arbeit stehen automatisierte Themenanalysen im Vordergrund. Während der Arbeit an diesem Thema wurden geeignete Vorgehensweisen entwickelt, wie Verfahren des Text-Mining in der Praxis für Inhaltsanalysen eingesetzt werden können. Unter anderem wurden Beiträge zur Visualisierung und einfachen Benutzung unterschiedlicher Verfahren geleistet (Niekler u. a., 2012). Weitere Beiträge entstanden im Rahmen der methodologischen Einordnung der computergestützten Inhaltsanalyse (Heyer u. a., 2014; Lemke u. a., 2015) und in der Definition innovativer Anwendungen im Bereich der computergestützten Inhaltsanalyse (Wiedemann u. Niekler, 2014; Niekler u. a., 2014a).

Die wissenschaftliche Untersuchung, die in dieser Arbeit durchgeführt wird, basiert auf eigenen Software-Implementierungen. Eine schematische Übersicht über die benötigten Softwarekomponenten und Ressourcen ist in dieser Arbeit in Anhang A dokumentiert. Die Verarbeitung, Verwaltung, Analyse und Visualisierung digitaler Textsammlungen steht dabei im Mittelpunkt. Experimente und Untersuchungen wurden komplett in dieser eigens entwickelten Software durchgeführt, die in anderen Projekten erfolgreich eingesetzt wird. Dabei entstand vorerst ein Prototyp (Niekler u. a., 2012), der später zu einer ausgereiften Softwareumgebung (Niekler u. a., 2014b) weiterentwickelt wurde und wird. Den Grundstock bildet dabei das Apache Framework UIMA (uima), welches die Funktionalitäten für die Analyse unstrukturierter Textdaten bereitstellt. Um dieses System herum wurden Verarbeitungsketten, Datenhaltung, Visualisierung, grafische Oberflächen, Möglichkeiten der Dateninteraktion, maschinelle Lernverfahren und Komponenten für das Dokumentretrieval implementiert. Dadurch werden die komplexen Methoden und Verfahren für die automatische Inhaltsanalyse einfach anwendbar und sind für künftige Projekte und Analysen benutzerfreundlich verfügbar. Sozialwissenschaftler, Politikwissenschaftler oder Kommunikationswissenschaftler können mit der Softwareumgebung arbeiten und Inhaltsanalysen durchführen, ohne die Details der Automatisierung und der Computerunterstützung durchdringen zu müssen.

Das diachrone Korpusformat nachrichtenorientierter Textquellen stellt besondere Anforderungen an die automatisierte Verarbeitung von Texten. Zum einen müssen die Texte so zugänglich gemacht werden, dass eine Analyse unter Beachtung der Zeitstempel möglich wird oder die Reihenfolge der Dokumente in einem Verarbei-

tungsprozess berücksichtigt wird. Zum anderen sind die Verfahren nicht für diese Art der Verarbeitung erdacht. Deshalb mussten für die Durchführung der Arbeit Anpassungen an bestehenden Verfahren vorgenommen werden, um die methodischen Anforderungen der Inhaltsanalyse erfüllen zu können. Dies betrifft insbesondere die Anwendung der Topic-Modelle. Diese sind nach der Literatur nicht auf verschiedene Untermengen anwendbar, wie es bei einer tagesweisen, monatsweisen oder jahresweisen Verarbeitung nötig wäre. Für diese Untermengen kann nur je ein Modell berechnet werden. Deshalb wurde die Möglichkeit ausgelotet, mehrere Modelle miteinander zu verknüpfen, um sequenzielle Analysen unterschiedlicher Dokumentmengen mit Topic-Modellen möglich zu machen (Niekler u. Jähnichen, 2012). Das Ergebnis dieser Arbeit ist in Abschnitt 3.2.4 ausführlich für die Anwendung in dieser Arbeit dokumentiert. In diesem Zuge wurde die Clustermethode aus dem „Topic Detection and Tracking“ für das Korpusformat angepasst. Weiterhin wurde das herkömmliche Verfahren für Kookkurrenzanalysen angepasst (Bordag, 2008; Biemann, 2012). Für die Anwendung in inhaltsanalytischen Aufgaben musste das Konzept der Nachbarschaftskookkurrenzen auf die Verwendung größerer Fenster erweitert werden, sodass die Analyse von gerichteten Kookkurrenzen möglich wird (Niekler u. a., 2014b). Diese alternative Vorgehensweise wurde durch den Autor implementiert und kommt unter anderem in Analysen und Aufsätzen zur Anwendung (Lemke u. Stulpe, 2015).

Kapitel 2

Technische und theoretische Grundlagen für die automatische Inhaltsanalyse von Themenstrukturen

Bei der analytischen Arbeit in Textarchiven werden verschiedene Aspekte dort manifestierten Wissens erforscht und zu beobachtet. Die empirische Kommunikationsforschung bietet mit dem Teilgebiet der Inhaltsanalyse eine umfassend untersuchte Methode, die die Beobachtung relevanter Aspekte in Inhalten erklärt und theoretisch fundiert.[1] Viele technische Implementierungen und Arbeiten zum Thema der EDV-gestützten Analyse von Textarchiven beziehen sich auf die Methoden der Inhaltsanalyse und nutzen sie. Meist nutzen diese Arbeiten einfache Zählverfahren. Beispielsweise werden Diktionäre herangezogen um automatisch festzustellen, wie oft und wo die darin enthaltenen Wortformen in einer Textquelle vorkommen (Stone, 1966). In West (2001) werden die computergestützten Methoden vielfältig für interessante Analysen eingesetzt. Dennoch ist deren Aussagekraft durch eine starke Reduktion von Bedeutungszusammenhängen und quantitative Auszählung eingeschränkt. Abgesehen von Scharkow (2012), der komplexere Methoden einführt, beschäftigen sich die Ausführungen also oft mit reiner stichwortgestützter Auszählung oder Methoden, die den inhaltlichen Kontext einzelner Wörter außer Acht lassen.

[1] Ausgezeichnete Einführungen in die Methode sind Merten (1995); Früh (2007); Krippendorff (2004); Neuendorf (2002).

In den folgenden Ausführungen wird in die Problematik der Methode der Inhaltsanalysen eingeführt. Als Grundlage der Erörterung dient die Anwendung in der Medien- und Kommunikationswissenschaft, die die theoretische und praktische Vorgehensweise der Inhaltsanalyse geprägt hat. Die Inhaltsanalyse bildet die methodische Grundlage der Arbeit und wird in einem eigenen Unterkapitel erläutert.

Da die Anforderungen und der Arbeitsaufwand einer manuellen bzw. herkömmlichen Inhaltsanalyse sehr komplex sein können, will diese Arbeit durch automatische, rechnerbasierte Verfahren beitragen, diese Aufgaben effizienter zu lösen. Dafür geeignete Verfahren werden in der automatischen Verarbeitung natürlicher Sprache, einem Teilbereich der Informatik, theoretisch und praktisch erforscht. In einem weiteren Teil dieses Kapitels werden diese Ansätze erarbeitet und diskutiert.

2.1 Inhaltsanalyse

Sozialwissenschaftliche oder kognitionswissenschaftliche empirische Untersuchungen stützen sich oft auf die Methode der Inhaltsanalyse, um gesellschaftliche Handlungen und Vorgänge auf Grundlage einer Hypothese zu untersuchen. Da viele Belegstellen für empirische Untersuchungen in gesprochener, niedergeschriebener oder beobachteter Form vorliegen, werden diese Inhalte untersucht. Das können gesprochene, gefilmte oder transkribierte Interviews, Nachrichtentexte sowie Kinofilme sein. Besonders interessant sind dabei Nachrichtentexte, da diese das öffentliche Interesse und die öffentliche Aufmerksamkeit von Themen wiedergeben. Der Fokus dieser Arbeit liegt auf Textanalysen, da redaktionelle Texte die Datengrundlage darstellen. Die inhaltsanalytische Arbeit umfasst verschiedene Ausprägungen wie strukturelle Analysen bzw. Schlüsse, semantische Analysen bzw. Schlüsse und pragmatische Analysen bzw. Schlüsse (Merten, 1995, vgl. S. 119ff.). Aus strukturellen Analysen lassen sich beispielsweise Zusammenhänge und Erkenntnisse bezüglich der Sprachkomplexität herstellen. Semantische Analysen benutzen Kontexte, die durch Klassifikation oder Kategorisierung hergestellt werden können und stellen eine der meistverwendeten Inhaltsanalysen dar, die Themenanalyse. Medienanalysen, Trendanalysen, Vergleiche und der Schluss auf die soziale Wirklichkeit werden dadurch ermöglicht.

Der Inhaltsanalyse geht es um die Erhebung empirischer Daten, die aus materialisierter Kommunikation gewonnen werden können. Dabei wird die Zielsetzung einer solchen Erhebung unterschiedlich definiert. Nach Krippendorff (2004, S. 18) gilt für die Inhaltsanalyse, dass „Inhaltsanalyse eine Forschungsmethode ist, um reproduzierbare und valide Inferenzen aus Texten [. . .]auf ihren Verwendungskontext zu ziehen.“ Weiterhin gilt nach Berelson (1984, S. 18), „Content analysis is a research techqi-

que for the objective, systematic and quatitative description of the manifest content of communication.“ Dabei müssen Inhaltsanalysen solide (reliable) bzw. verlässlich, reproduzierbar (replicable) und gültig (valid) sein. Dies bedeutet, dass Inhaltsanalysen, die mit gleichen Daten und Methoden arbeiten, zum selben Ergebnis kommen müssen. Zusätzlich müssen die Ergebnisse gegen andere empirische Befunde standhalten können. Die Definitionen spiegeln verschiedene Sichtweisen auf Inhaltsanalysen wieder. Während Krippendorff (2004) und Berelson (1984); Früh (2007) den deskriptiven Charakter der Inhaltsanalysen hervorheben, spricht Merten (1995) von einer Inferenz auf die soziale Wirklichkeit. Neben einer Beschreibung der Inhalte kann die Form der Aussage (Lesbarkeit, Stil), Beziehungen zwischen Inhalt und Kommunikator (Intention, Persönlichkeit), Beziehungen zwischen Inhalt und Rezipient und die Wirkung (Aufmerksamkeit, Einstellung, Verhalten) untersucht werden (Merten, 1995, vgl. S. 55). In Rössler (2005, vgl. S. 22 f.) wird darauf hingewiesen, dass die Rückschlüsse auf die Wirklichkeit, die nach der Auswertung der manifesten Materialien getroffen werden können, das eigentliche Potential einer Inhaltsanalyse bilden, anstatt die Inhalte lediglich zu beschreiben. Die Interpretationsfähigkeit der Analyseergebnisse ist demnach ein wichtiges Kriterium für die Durchführung einer Inhaltsanalyse. Auf die Interpretationsfähigkeit der Analysen bezieht sich oft die Kritik, die an Inhaltsanalysen geübt wird. Denn oft muss die „[. . .]Quantifizierung notwendig selektiv vorgehen [. . .]und damit möglicherweise wesentliche Beziehungen, die nur interpretativ sind, also aus der Textqualität des Textes zu erschließen sind [. . .]“ aussparen (Merten, 1995, S. 98).

2.1.1 Methodik und Eigenschaften

Die folgenden kontrastiven Unterscheidungen und Merkmale von Inhaltsanalysen geben Aufschluss über die Eigenschaften von Inhaltsanalysen und weisen auf Limitierungen hin, die sie mitbringen. Durch die vorhandenen Unterscheidungen muss jedes automatische Verfahren nach den definierten Kriterien bewertet werden. Nur so kann eine definierte Werkzeugpalette automatisierter Methoden für die Inhaltsanalyse erstellt werden.

Qualitative und Quantitative Inhaltsanalysen

In der wissenschaftlichen Debatte innerhalb der Inhaltsanalyse gibt es durch die Kritikpunkte an der Interpretationsfähihkeit der Analysen eine „Qualitativ-quantitativ-Debatte“ (Früh, 2007, vgl. S. 67 ff.), deren Inhalt hier nicht nachvollzogen werden soll. Aus dieser Debatte entstehen allerdings Konsequenzen für die Einordnung und

Verwendung der Inhaltsanalyse und deren Ergebnisse. In der Inhaltsanalyse werden demnach qualitative und quantitative Messverfahren unterschieden. Während quantitative Messverfahren eine Analyse in viele zu messende quantitative Variablen zerlegen, die in großen Umfängen gemessen werden können, so konzentrieren sich die qualitativen Verfahren auf wenige Einzelfälle und wenige beobachtbare Variablen (Früh, 2007, vgl. S.70). Weiterhin wird bei der Interpretation von Analyseergebnissen zwischen quantitativen Aussagen und qualitativen Aussagen unterschieden. Während quantitative Aussagen eine tatsächlich gemessene Größe wiedergeben also deskriptiv wirken, können qualitative Aussagen aus einem Inferenzschluss von mehreren quantitativ gemessenen Variablen gebildet werden (Früh, 2007, vgl. S. 67). Um messbare Variablen für die Quantifizierung zu bilden, müssen bestimmte Inhalte als Kategorie zusammengefasst werden. Dies geschieht per Definition und theoretischer Vorarbeit. Es ist also eine qualitative Vorarbeit zur quantitativen Messung notwendig. Die Quantitäten haben demnach eine Bedeutung, wie Früh (2007, vgl. S. 67 ff.) anmerkt. Die Merkmalsdimensionen einer Kategorie müssen im Gesamtzusammenhang festgelegt werden und empirisch messbar sein (Merten, 1995, vgl. S. 98). Anhand dieser Festlegungen ist es möglich, ein Ordnungssystem aufgrund der Nominalskala zu schaffen, auf das der Inhalt verteilt werden kann (Merten, 1995, vgl. S. 98) und innerhalb derer weitere Messungen angestellt werden können. Hier ergibt sich für die Arbeit die erste Anforderung. Die untersuchten Verfahren müssen in der Lage sein, Verwendungszusammenhänge in Textdaten zu erfassen und die Grundlage für die EDV-gestützte Umwandlung der Nominalskala in ein Ordnungssystem bilden. Dies ist zumindest ein Hinweis auf nicht überwachte Verfahren, deren Funktion keine externen Trainingsdaten benötigt, und deren Potential explorative Suchen in großen Textmengen ermöglicht. Ist der Interpretationsprozess einer quantitativen Bestimmung von Kategoriezuordnungen nachgestellt, so müssen die Kategorien so beschaffen sein, dass ein Schluss auf die Forschungsfrage möglich sein kann.

Aus dieser Diskussion der Begriffe ergeben sich folgende Konsequenzen für die Möglichkeiten einer computergestützten Inhaltsanalyse:

1. Es kann keine qualitative Messung mit elektronischen Verfahren im Sinne der qualitativen Beurteilung von Variablen vorgenommen werden, da diese immer vorhandene menschliche Interpretation voraussetzt.

2. Durch die Eigenschaften quantitativer Messungen ist es möglich, das Potential von Text-Mining-Verfahren auszunutzen, um diese effizienter zu gestalten.

3. Es ist möglich, falls eine geeignete Operationalisierung zu messender Variablen durch das Text-Mining existiert, qualitative Aussagen mit Hilfe automatisierter Verfahren zu treffen, sofern durch Menschen interpretierbare Ergebnisse erzeugt werden.

Deskription und Inferenz

Aus der Diskussion über quantitative und qualitative Messungen ergeben sich Fragen, welche Aussagekraft die erhobenen Daten haben. Dabei können zwei Verwertungsaspekte hervorgehoben werden. Einerseits werden Inhaltsanalysen zur reinen Deskription der Inhalte durchgeführt werden, was gleichzeitig die Grundlage einer jeden Inhaltsanalyse ist. Dabei werden syntaktische Merkmale, inhaltliche Kategorien, Akteure, Quellen, Nachrichtenwerte, Werbezeiten oder jugendgefährdende Inhalte identifiziert (Merten (1995, vgl. S. 23), Rössler (2005, vgl. S. 25)). Andererseits versucht die Inferenz eine Korrelation von Text-internen Merkmalen und Text-externen Merkmausausprägungen herzustellen (Merten, 1995, vgl. S. 23). Dabei werden drei Inferenzschlüsse unterschieden. Den Schluss auf den Rezipienten (z.B. Verständlichkeit), den Kommunikator (z.B. Stil) und die Situation (z.B. Psychologie), in der die Kommunikation stattgefunden hat. In diesem Sinne ist die inhaltsanalytische Inferenz nicht gleichbedeutend mit der statistischen Inferenz, bei der von Eigenschaften der Textstichprobe auf Eigenschaften der Grundgesamtheit geschlossen wird (Merten, 1995, vgl. S. 110). Die Interpretation der Inhaltsanalyse unterliegt weiteren theoretischen Annahmen. So kann nach dem Repräsentationsmodell interpretiert werden oder nach dem Instrumentalmodell. Das Repräsentationsmodell geht davon aus, dass die Inhalte auch die soziale Wirklichkeit spiegeln. Das Instrumentalmodell geht von einer bewussten Verzerrung der Inhalte durch die Wiedergabe in den Medien aus, sodass bei einer Gültigkeit dieses Modells davon ausgegangen werden muss, dass nicht die direkte soziale Wirklichkeit inferiert wird, sondern die Medien selbst analysiert werden. Dabei können Fragen nach den Absichten oder Ursachen der Verzerrung ebenso wichtig sein. Damit ergeben sich mehrere Arten der Messung wie

- die Abbildung der Wirklichkeit (Repräsentationsmodell),
- die Medienanalyse (Instrumentalmodell) und
- die Trendanalyse.

Der letzte Punkt stellt insbesondere eine zentrale Funktion von Inhaltsanalysen dar, da bei einer angenommenen Stabilität Veränderungen am besten festgestellt werden

können. Je nachdem, ob soziale Wirklichkeit oder Medien analysiert werden kann die Transanalyse jeweilige Veränderungen darstellen. Somit kann eine Inferenz eben absolut (Zustand) oder relational (Veränderung) ausgedrückt werden.

Im wissenschaftlichen Bereich ist die reine Deskription oft nicht sinnvoll, sodass die inhaltsanalytische Inferenz anhand der gegebenen Messdaten immer durchführbar sein muss. Hinsichtlich einer Operationalisierung durch linguistische Verfahren muss also immer geprüft werden, ob die deskriptiven Variablen immer gültige Aussagen für die Zustandsbeschreibung einer Situation liefern. Merten sieht Probleme, valide Inferenzschlüsse aus linguistischen Verfahren zu treffen (Merten, 1995, vgl. S. 119). Ein wichtiges Anliegen der Arbeit muss deshalb in der Anwendung von Verfahren liegen, die valide Inferenzschlüsse auf die Wirklichkeit, die Medien und die darunter liegenden Trends zulassen.

Deduktiv und Induktiv

Nach Früh (2007, vgl. S. 147) ist die Inhaltsanalyse eine „[. . .]offengelegte, systematische Suchstrategie“. Dabei kann entweder der bekannte Zusammenhang mehrerer Variablen überprüft und gegebenenfalls gestärkt oder falsifiziert werden. Dies entspricht einem deduktiven Vorgehen. Es kann ein unbekannter Zusammenhang hergestellt werden, was einem induktiven Vorgehen entspricht. Deshalb wird die deduktive Suche zur Überprüfung von Hypothesen durchgeführt und die induktive Suche, um Hypothesen und Forschungsfragen aus gegebenen Daten abzuleiten. Dies hat insbesondere für die Beurteilung computergestützter Verfahren Relevanz, da die Leistungsfähigkeit hinsichtlich der Möglichkeiten zum deduktiven oder induktiven Arbeiten beurteilt werden müssen.

2.1.2 Planung, Struktur und Ablauf

Am Anfang einer Inhaltsanalyse stehen zweifelsfrei Fragen und Hypothesen. Aufbauend auf einer deduktiven oder induktiven Suchstrategie wird der Untersuchungsgegenstand festgelegt. Nach Früh (2007, vgl. S. 102) können vier Phasen einer Inhaltsanalyse definiert werden. Es werden die Planung, die Entwicklung, der Test und die Anwendung der Inhaltsanalyse unterschieden. Die Planungsphase dient zur Entwicklung von Hypothesen anhand eines konkreten Problems, was einer intellektuellen und qualitativen Leistung entspricht. Im Gegensatz dazu muss in der Entwicklungsphase die Untersuchung der Hypothesen konkret operationalisiert werden. Diese Operationalisierungen müssen anhand von kleinen Stichproben auf der zu bearbeitenden Textmenge getestet und intellektuell bewertet werden. Anhand dieser Bewertung

müssen gegebenenfalls Anpassungen an Theorie und Operationalisierung vorgenommen werden. Ergibt ein solcher Pre-Test valide und reliable Ergebnisse, so wird in einer Analyse auf der gesamten Untersuchungsmenge die eigentliche Inhaltsanalyse durchgeführt.

Die Hypothesenbildung hängt stark von einer verwendeten oder referenzierten Theorie ab. In Inhaltsanalysen werden oft Kommunikations-, Diskurs-, System- oder Medientheorien zugrunde gelegt. Für die Operationalisierung der Problemstellung wird in der Entwicklungsphase meist ein Kategoriensystem entwickelt.

Wichtige Begriffe der Inhaltsanalyse

Im Vorfeld einer Analyse müssen verschiedene Rahmenbedingungen geklärt werden. Automatische Analysen müssen sich mit diesen Bedingungen auseinander setzen und in einen Zusammenhang mit diesen Bedingungen gebracht werden. Aus diesem Grund werden an dieser Stelle wichtige Begriffe und Methoden der Inhaltsanalyse angesprochen, damit in der späteren Argumentation darauf Bezug genommen werden kann. Die hier gezeigten Begriffe sind teil der Vorbereitung und Auswertung einer Inhaltsanalyse und müssen bei der Untersuchung automatischer Algorithmen beachtet werden.

Operationalisierung: Die Operationalisierung ist die konkrete Formulierung von Methoden und Vorgehen, die direkt von der Forschungsfrage abgeleitet wird und diese valide und reliabel beantworten muss (Bonfadelli, 2002, vgl. S. 87). Viele Untersuchungen operationalisieren die Fragestellung als Kategoriensystem für die quantitative Einordnung der Inhalte.

Reliabilität: Unter der Reliabilität sind verschiedene Gütekriterien von Inhaltsanalysen zusammengefasst. Im Vordergrund steht hierbei die Forderung, dass die Ergebnisse einer Inhaltsanalyse bei einer wiederholten Messung reproduzierbar sind (Rössler, 2005, vgl. S. 183).

Validität: Die Validität bestimmt, ob die Maße, Variablen und definierten Kategorien geeignet sind, um die Forschungsfrage hinter der Inhaltsanalyse zu klären (Früh, 2007, vgl. S. 196). Dabei geht es im Wesentlichen um die interpretierbare Bedeutung der Messungen. Die Vollständigkeit der zu messenden Phänomene, vor allem die Vollständigkeit von Kategorien und deren Ausprägungen, wird dabei unter der Inhaltsvalidität zusammengefasst. **Skalen:** Für die Erfassung und Messung der Inhalte in quantitativen Inhaltsanalysen können verschiedene Skalen verwendet werden, um die Inhalte zu kodieren. Die Anwendung unterschiedlicher Skalen ist nötig, da nicht alle Inhalte gleichmäßig erfasst werden können. Dabei wird unterschieden, ob

die Inhalte einer Gruppe bestimmter Eigenschaften (Nominalskala), einer Sortierung von Gruppen oder Werten(Ordinalskala, Intervallskala) oder einem konkreten Wert (Rationalskala) zugeordnet werden (Merten (1995, vgl. S. 96), Früh (2007, vgl. S. 32), Krippendorff (2004, vgl. S. 161 ff.)). **Grundgesamtheit:** Als Grundgesamtheit wird die Gesamtheit des zur Verfügung stehenden Materials bezeichnet, welches alle möglichen Fälle und kategorialen Ausprägungen der Analyse enthält. Die Definition der Grundgesamtheit leitet sich direkt aus der Fragestellung ab und trennt relevante von nicht relevanten Quellen.

Sampling- und Auswahleinheit: In der Bestimmung der Auswahleinheit wird entschieden, welche Materialien in einer Analyse verwendet werden. Die verwendeten Materialien, wie zum Beispiel Artikel aus Tageszeitungen, werden aus einer Fülle an Material, die Gesamtheit aller Tageszeitungen, gewählt und als Zufallsstichprobe aus dieser Gesamtheit gezogen (Rössler (2005, vgl. S. 39), Krippendorff (2004, vgl. S. 98), Merten (1995, vgl. S. 81)). Die Strategie hinter der Stichprobenziehung kann unterschiedlich sein.[2] Unterschieden wird zwischen willkürlicher Auswahl, einer bewussten Auswahl und einer Auswahl die auf der Wahrscheinlichkeitstheorie beruht. Beispielsweise können bewusst typische Fälle gewählt werden, wenn diese schon bekannt sind oder die Auswahl wird nach einer angenommen Wahrscheinlichkeitsverteilung der Grundgesamtheit durchgeführt.

Analyse- und Kodiereinheit: Diese Einheiten bezeichnen die Elemente, über die eine Aussage getroffen werden soll. Kodiereinheiten wird genau ein Code zugeordnet, was bedeutet, dass die Kodiereinheit genau einer Kategorie zuordnet wird. In Texten können Kodiereinheiten demnach Wörter, Sätze, Personennamen oder ganze Artikel sein, wenn Artikelthemen zuordnet werden sollen (Rössler (2005, vgl. S. 40), Früh (2007, vgl. S. 95), Krippendorff (2004, vgl. S. 99), Merten (1995, vgl. S. 281)). Oft werden die Analyse- und Kodiereinheit gleichgesetzt, was aber nicht immer der Fall ist (Früh, 2007, vgl. S. 95). So können Wörter in einem Artikel beispielsweise einzeln codiert werden, als Kodiereinheit, und so verrechnet werden, dass der kategoriale Anteil auf die Analyseeinheit Artikel codiert wird. Diese Unterscheidung spielt insbesondere eine Rolle, wenn die Kodierung durch computergestützte Verfahren wortbasiert funktioniert aber eine Entscheidung über die Zugehörigkeit von Sätzen, Absätzen oder ganzen Dokumenten erfolgen soll. In diesem Fall muss es eine sinnvolle Strategie geben, um viele Einzelkodierungen zu abstrahieren oder zusammenzufassen.

[2] Die Strategien sind in Krippendorff (2004, vgl. S. 114 ff.) und Merten (1995, vgl. S. 283) umfassend beschrieben.

Kontexteinheit: Um einer Analyseeinheit eine Kodierung zuzuordnen bedarf es in einigen Fällen mehr, als nur die Kodiereinheit bzw. die Analyseeinheit auszuwerten. Oft muss der Kontext, in dem eine Analyseeinheit bewertet werden darf, definiert werden (Krippendorff (2004, vgl. S. 101), Rössler (2005, vgl. S. 42), Merten (1995, vgl. S. 282)). Der Kontext kann dabei als syntaktische Einheit, wie zum Beispiel ein Absatz, definiert sein oder als Fläche, Länge oder Abstand zur Analyseeinheit oder über externe Referenzen, die zur Vergabe der Codes herangezogen werden.

Entdeckungszusammenhang: Bei einer Inhaltsanalyse wird zunächst gefragt, welche Annahmen und Theorien empirisch überprüft werden sollen (Merten, 1995, vgl. S. 314). Dabei wird geklärt, welche Theorien infrage kommen oder welche Probleme untersucht werden. Das Ziel der Untersuchung muss hinreichend definiert werden. Dieser Prozess wird als Entdeckungszusammenhang bezeichnet.

Begründungszusammenhang: Um das Ziel der Analyse zu erreichen, muss die Inhaltsanalyse entworfen und geplant werden. Dies geschieht grob in den Phasen Planung, Entwicklung, Test und Anwendung. Die gesamte Durchführung der Inhaltsanalyse wird als Begründungszusammenhang bezeichnet (Merten, 1995, vgl. S. 314). Der Begründungszusammenhang muss schlüssig und intersubjektiv nachvollziehbar dokumentiert werden, sodass die Analyse reproduzierbar und verständlich ist. Die Dokumentation kann sehr individuell und dadurch für andere schwer nachvollziehbar sein. Beispielsweise müssen Mehrdeutigkeiten in der Interpretation vermieden werden. Hier bieten computergestützte Verfahren das Potential, zumindest die Erstellung der Daten sehr transparent dokumentieren zu können.

Methodik der Kategorienbildung bei Inhaltsanalysen

Um theoretisch definierte oder empirisch festgelegte Zusammenhänge zu messen, werden Empirie- oder theoriegeleitet Kategoriensysteme gebildet, innerhalb derer Variablen für die Überprüfung der Hypothesen definiert werden können (Merten, 1995, vgl. S. 198 f.). Kategorien können in diesem Sinne formal, inhaltlich oder wertend sein (Rössler, 2005, vgl. S. 104 ff.). Formale Kategorien können aus quantitativen Variablen wie Textmenge, Seitenzahl oder Präsentationsgröße im Medium bestehen. Diese Kategorien sind jedoch rein deskriptiv und lassen nur wenige Inferenzen, wie zum Beispiel Wichtigkeit des Textes in einem Medium, zu. Die Entwicklung inhaltlicher Kategorien ist die konkrete trennscharfe Unterteilung der Textinhalte in definierte Einheiten. Die Definition der Kategorien legt demnach die Dimension und den Umfang der Daten fest, die benötigt werden (Früh, 2007, vgl. S. 81). Die Entwicklung inhaltlicher Kategorien soll die Inhalte der Texte klassifizieren. Diese Kategorien

können auf Grundlage von Themen, Ereignissen, Akteuren, Handlungsträgern oder einem Aktualitätsbezug definiert werden. Weiterhin können Kategorien eingeführt werden, bei denen mehrere Indikatoren im Inhalt vorgefunden werden müssen.

Bei der Bildung der Kategorien ist darauf zu achten, dass die zu messenden Kategorien folgende Kriterien erfüllen, damit eine Kodierung und Zuordnung zuverlässig ist (Merten (1995, vgl. S. 98 f.), Früh (2007, vgl. S. 86), Bonfadelli (2002, vgl. S. 90),Neuendorf (2002, vgl. S. 119)):

- Relevanz/ Erschöpfend: Die Kategorie soll für die Fragestellung geeignet sein und gerade so komplex sein wie nötig.
- Unabhängigkeit: Die Kategorien dürfen nicht miteinander korrelieren.
- Vollständigkeit: Das Kategoriensystem muss alle zu untersuchenden Inhalte erfassen und abbilden können. Falls eine eingeschränkte Kategorisierung von Interesse ist, so muss immer eine offene Kategorisierung mit einer Residualkategorie (offenes Kategoriensystem) gewählt werden, sodass nicht passende Inhalte in eine allgemeine, offene Kategorie einsortiert werden können.
- Eindimensionalität: Die Kategorien sollen jeweils nur einen Aspekt messen und klassifizieren, wie z.B. Thema, Bewertung oder Akteure. Eine Kategorie darf nicht verschiedene Prinzipien zur Klassifikation nutzen.
- Trennschärfe: Die Kategorien sollen in ihrem Inhalt eindeutig abgrenzbar sein.

Die Inhaltsanalyse nach einem definierten Kategoriensystem kann nach zwei unterschiedlichen Ansätzen erfolgen. So kann einerseits eine im Vorfeld bekannte oder gebildete Theorie die Grundlage des Kategoriensystems bilden, mit dessen Hilfe die Fragestellung bearbeitet wird. Es wird versucht, theoretische Konstrukte in ein Kategoriensystem zu überführen, sodass festgelegte Hypothesen durch eine deduktive Suche in den Textinhalten überprüft werden können. Ist für eine Fragestellung bereits ein Kategoriensystem entworfen, so kann es auch auf neue Daten angewendet werden, wenn die Frage auf der Grundlage von anderen Daten neu beantwortet werden soll. Andererseits müssen Kategoriensysteme auch gebildet werden, wenn die Untersuchung an einem unbekanntem Zusammenhang erfolgt. Die Theoriebildung ist nicht vollständig abgeschlossen und kann nicht zur Definition der Konzepte und Kategorien herangezogen werden. In einem solchen Fall muss eine induktive Suche in den Daten vorgenommen werden. Das bedeutet, dass Kategorien oder zugrunde liegende Konzepte anhand der Daten abgeleitet und erstellt werden müssen. Die

als empirische Kategorienbildung bezeichnete Aufgabe muss intersubjektiv nachvollziehbar sein, sodass die Kategorien und Schlüsse einer Analyse im Nachhinein zu verstehen sind. Das heißt, dass auch die empirische Kategorienbildung einem Formalismus folgen muss, der eine Nachvollziehbarkeit ermöglicht. Die Kritik an diesem Verfahren liegt in der Gefahr, dass unvollständige Kategoriensysteme gebildet werden, da nicht alle Daten gesichtet werden können. Allerdings stehen moderne computergestützte Verfahren diesem Argument entgegen, da sie gerade für die Verarbeitung vieler Dokumente erdacht sind. Der sogenannte „Distant Reading“ Ansatz von Moretti spiegelt diese technische Möglichkeit wieder (Moretti, 2005, vgl. S. 1). Da mit computergestützten Methoden die Limitierung auf wenige Dokumente nicht existiert, bekommt die empirische Kategorienbildung eine neue Relevanz. Sie ist immer eine subjektive und selektive Sichtweise, die auf dem Erfahrungsschatz und dem Wissen der Analysten basiert (Bilandzic u. a., 2001). Bilandzic u. a. (2001) schlägt vor, dass die Inhalte heuristisch-theoretisch segmentiert und innerhalb der gebildeten Segmente Zusammenfassungen gebildet werden. Der Entdeckungszusammenhang wird damit nachvollziehbar. Die Inhalte in den Segmenten, die Sätze, Absätze oder ganze Dokumente sein können werden paraphrasiert und strukturiert. In drei Schritten werden innerhalb dieser Strukturen verallgemeinerbare Segmente als Kategorien abgeleitet. Der kreative Prozess der Kategorie- oder Theoriebildung wird dadurch systematisch und transparent und die individuelle Entscheidung des Forschers wird nachvollziehbar. Ein weiterer Beitrag zur empirischen Kategorienbildung ist in dem Band Wirth u. Lauf (2001) erschienen. Dort wird aufgezeigt, dass die Eingrenzung auf Segmente noch nicht ausreicht, um die Kategorienbildung vollständig transparent zu gestalten. Früh (2001) bezieht sich auf die Möglichkeit innerhalb der definierten Kodiereinheiten Propositionen zu extrahieren, sodass der objektive Bezug der Inhalte formalisiert erfasst werden kann. Der Abstraktionsgrad der Propositionen oder die Art und Weise sie zu verdichten oder zu generalisieren richtet sich nach der eigentlichen Fragestellung. Im ersten Schritt wird ein Formalismus genutzt, um nachvollziehbar Zusammenhänge und Bezüge zu extrahieren. Dies verursacht sehr kleinteilige Informationen und es muss je nach Forschungsfrage stark abstrahiert werden, um ein Kernthema zu finden. Dennoch ist festzuhalten, dass die Propositionen als Orientierung dienen und für eine weitere Bestimmung der Kategorieneigenschaften eingesetzt werden können. An den Überlegungen zur empirischen Kategorienbildung kann allgemein gesagt werden, dass

- in einem ersten Schritt eine Kodiereinheit selektiert werden muss innerhalb derer Kodierungen vorgenommen werden,

- darin weitere formale Einheiten bestimmt werden müssen, sodass dort nach unterschiedlichen Nennungen und Ausprägungen bezüglich einer Fragestellung gesucht werden kann und
- alle Ausprägungen, seien es Propositionen oder Zusammenfassungen des Inhalts, zu einer Klassifizierung der Texte zusammengefasst werden können.

In Anlehnung an Früh (2007, vgl. S. 157) ergibt sich für die empirische Kategorienbildung folgendes Vorgehen:

- Selektion / Reduktion: Die Auswahl relevanter Textstellen im Text.
- Bündelung: Die Gruppierung der Textstellen nach Gemeinsamkeiten.
- Generalisierung / Abstraktion / Bezeichnung: Die Zuweisung von Labels und gemeinsamen Bedeutungsinhalt von Textstellen.
- Rückbezug auf Theorie: Anwendung der als relevant betrachteten Textstellen zur Generierung neuer Hypothesen, die über empirische Kategorienbildung untersucht werden können.

Die Kritik, die an derartigen empirischen Vorgehensweisen geübt wird, begründet sich auf der Komplexität dieser Verfahren. So können durch die Komplexität und durch manuelle Arbeit nicht alle möglichen Ausprägungen untersucht werden. Daraus kann eine nicht ausreichende Klassifikation resultieren (Bilandzic u. a. (2001), Früh (2007, vgl. S. 163)). Esist möglich, dass wichtige Kategorien oder Ausprägungen nicht gesehen werden und durch diese Unvollständigkeit keine gültigen Verallgemeinerungen getroffen werden können. Weiterhin gilt, dass die Definition und Komplexität der Kategorien komplett vom Forscher übernommen wird (Früh, 2007, vgl. S. 162), was einerseits schwer nachvollziehbar ist und andererseits immer von der individuellen Erfahrung abhängt, die eine verzerrende Sichtweise in das Kategoriensystem induzieren kann. Aus diesen Gründen muss die empirische Kategorienbildung formal, nachvollziehbar und repräsentativ durchgeführt werden. Durch eine möglichst genaue Kategorisierung soll der Interpretationspielraum eingeschränkt werden. In zeitabhängigen Medien oder über die Zeit beobachteten Medien können Kategorien aber auch entstehen oder wieder verschwinden, was die Möglichkeiten wiederverwendbarer Kategorien einschränkt (Merten, 1995, vgl. S. 101). Aus diesem Grund ist es umso wichtiger Mittel und Wege zu finden, die empirische Kategorienfindung zu stützen und weiter zu entwickeln, da die Datenmengen immer größer und deren Inhalt immer dynamischer und schnelllebiger wird.

In diesem Sinne muss die Erfassung und Bearbeitung neuer Kontexte und Thematisierungen und deren interne Dynamik in immer größeren Textkollektionen stattfinden. Aus diesen Erörterungen ergibt sich für diese Arbeit die Konsequenz, dass die untersuchten Verfahren einerseits ihren Beitrag zur empirischen Kategorienbildung zeigen müssen und andererseits Kategoriensysteme erzeugen, die den oben genannten Anforderungen entsprechen.

2.1.3 Themenanalysen

Die inhaltsanalytische Untersuchung von Texten auf deren Themengehalt ist eine der meist verwendeten Analyseformen. In vielerlei Hinsicht erlaubt diese Form der Untersuchung eine Beurteilung verschiedener Dimensionen von Texten. In Tageszeitungen kann beispielsweise der thematische Gehalt verschiedener Publikation analysiert werden, um deren inhaltlichen Fokus zu vergleichen. Im Allgemeinen können Themenvergleiche über die Zeit Aufschluss darüber geben, wie sich das mediale oder literarische Interesse für bestimmte Themen zeitlich verändert. In der Frage, was unter einem Thema zu verstehen ist, existieren viele unterschiedliche Ansätze und meist wird die kategoriale Unterscheidung von Themen anhand einer konkreten Fragestellung definiert. Dennoch muss bei der Definition von Themenkategorien auf Grundlagen und Theorien zurückgegriffen werden, die es erlauben eine untersuchte Kategorie auch als Thema zu bezeichnen. Während sich die Methode der Inhaltsanalyse selbst eine offene Themendefinition gibt, die sehr allgemein gehalten wird und darauf ausgelegt ist Informationen zu reduzieren und messbare Größen zu definieren (Roberts (1997, Vgl S. 37), Merten (1995, vgl. S. 147)), so ergeben sich dennoch Probleme mit einer validen Konzeption von thematischen Kategorien. Einerseits können Themen ereignisbezogen sein. Andererseits können Themen gesellschaftliche Diskurse darstellen, in die das Geschehen eingebettet ist. Diese sind wiederum Bestandteil von gesellschaftlichen Feldern (Rössler, 2005, vgl. S. 122). Diese Einordnung weist darauf hin, dass es in diesem Themenverständnis eine hierarchische Ordnung und eine gewisse Granularität bzw. Auflösung gibt, in denen Themen betrachtet werden. Dennoch hilft diese Einordnung nicht, herauszufinden, welche Textinhalte für eine Identifikation des Textthemas hilfreich sein können. Die genannte Unterscheidung gibt lediglich Auskunft über die Existenz solcher inhaltlicher Zusammenhänge. Sie zu finden und zu strukturieren bedarf einer weiteren Betrachtung des Themenbegriffs. Für die inhaltliche Definition von Themen finden sich mehrere Sichtweisen, die vornehmlich aus der sprachwissenschaftlichen Auseinandersetzung stammen. Im Gegensatz zu linguistischen Betrachtungen, die überwiegend aufgrund der grammati-

schen Tiefenstruktur von Sätzen und Texten basieren, existieren Definitionsansätze, die lediglich eine gewisse inhaltliche Fokussierung hinweisen, aber keine Aussage darüber machen, wie dieser Fokus aus einem Text extrahiert werden kann. Letztlich gibt es weitere Definitionen aus der Alltagssprache, die aber, da sie einem Allgemeinverständnis entsprechen, dennoch herangezogen werden können. An dieser Stelle sollen die angesprochenen Ansätze genauer untersucht werden, damit die Einteilung und Leistungsfähigkeit der automatisierten Themenextraktion anhand der existierenden Definitionen und Theorien beurteilt werden kann.

Die alltagssprachliche Definition des Themenbegriffs findet sich unter anderem im Duden (Dudenredaktion (Bibliographisches Institut), 2004, vgl. S. 963), wo das Thema als Gegenstand, Gesprächsstoff oder Leitgedanke geführt wird. Weitere synonym verwendete Worte und Hinweise auf das Thema als Angelegenheit, Anliegen, Aufgabe, Causa, Fall, Fragestellung oder Problem finden sich in den in Online Enzyklopädien (Wiki 4; DWDS) oder dem Projekt Deutscher Wortschatz, welches Wortverwendungszusammenhänge durch die distributionale Semantik großer Textsammlungen, vor allem Nachrichtentexten, bestimmt und damit einen guten Überblick über alltagssprachliche Gebräuche liefert (wortschatz 2). Für den Begriff „Thema“ liefert das Portal, in Form von Satzkookkurrenzen, den Hinweis, dass ein Thema wichtig, heikel oder zentral sein kann und es wird darüber diskutiert oder gelesen.[3] Eine dieses Verständnis treffende Definition findet sich in Lötscher (1987, S. 58) und er schreibt, das „Thema ist ein Gegenstand, der in einem Text (zentral) zur Sprache kommt.“ Diese Definition schließt ein, dass eine Thema in mehreren Texten zur Sprache kommen kann und den Mittelpunkt der Texte bildet. Einige Definitionen der Linguistik erlauben die Erweiterung auf mehrere Texte nämlich nicht und schränken die Verwendung eines solchen Themenbegriffs ein. In einer systemtheoretischen und soziologischen Sichtweise von Luhmann (1979, S. 13) kann man Themen als „[. . .], mehr oder weniger unbestimmte und entwicklungsfähige Sinnkomplexe verstehen, über die man reden und gleiche, aber auch verschiedene Meinungen haben kann [. . .].“ Die linguistischen Betrachtungen des Themenbegriffs gehen dagegen formaler und anhand von Texteigenschaften bei einer Definition vor. Diese Theorien beziehen sich in der Betrachtung des Themas meist auf Sätze und darin enthaltene Propositionen. Die Proposition ist die grundsätzliche Aussage oder der Sinn eines Satzes (Brinker, 1988, vgl. S. 14). Es gibt keine Bestimmung, ob sich dieser Aussagebegriff auf Objekte im Satz beziehen muss. Es ist jedoch ein Unterschied zur Illukation zu ziehen, welche die

[3] Kookkurrenzen messen die gemeinsame Verwendung von Wortformen in einer Analyseeinheit

Sprechintention beschreibt. So haben die Sätze „Franz spielt konzentriert.“, „Spielt Franz konzentriert?“ und „Franz, spiel konzentriert!“ dieselbe Proposition (Franz, konzentriert spielen) aber unterschiedliche Illukationen. Somit kann die Proposition bei unterschiedlicher Intention als die gemeinsame Aussage oder der gemeinsame Sinn, wie von Gottlob Frege (Frege, 1993, vgl. S. 35 ff.) erläutert, verstanden werden.

Auf dieses Konzept bezieht sich die Theorie der Makropropositionen (van Dijk, 1980). Es werden Regeln eingeführt, die es erlauben, die Propositionen in einem Text zu aggregieren und zu abstrahieren, sodass allgemeingültige Makropropositionen entstehen, die sich über mehrere Texte decken können. Diese Makropropositionen dienen dazu das Textthema anhand von Textmengen darzustellen, die eine globale Bedeutung haben und die Semantik eines Textes abbilden (van Dijk, 1980, vgl. S. 49). Er spricht verschiedene Abstraktionsgrade an, bei denen die Satzaussagen immer mehr verallgemeinert werden. Die angesprochenen Abstraktionsregeln Auslassen, Verallgemeinern und Konstruktion dienen der schrittweisen Informationsreduktion im Text, wie von der inhaltsanalytischen Theorie gefordert. Die Regeln sind allerdings sehr subjektiv anwendbar und hängen von den Fähigkeiten des Analysten ab, was deutlich wird wenn die Regel „Auslassen“ betrachtet wird, die erfordert, dass genau beurteilt werden kann, was wichtig ist und was nicht (Lötscher, 1987, vgl. S. 41). Beispiele die in der Literatur angeführt werden verstärken diesen Eindruck noch mehr, da keineswegs klar wird auf welcher Grundlage oder Abstraktion diese Regeln angewendet werden sollen (van Dijk, 1980, vgl. S. 47). Dennoch zeigt dieser Ansatz, dass es einen Themenkern in Form von abstrakten Propositionen gibt, um den das Thema entfaltet wird. Der Themenkern ist Kern- und Ausgangspunkt der „'Darstellung' [...], die gegenüber dem reinen Inhalt, dem 'Stoff', relativ selbstständig ist und [...]sogar mehreren Texten zugehören kann [...], die gewissermaßen ein ganzes Modell von objektiven Beziehungen und Prozessen als Wesen des Textes ausdrückt (Agricola, 1976).“ Er beschreibt, dass die Propositionen, oder in seinem Sinne Aktanten, immer wieder verwendet werden und mit anderen unbekannten Propositionen aber eben besonders auch untereinander in Beziehung treten. Das bedeutet, dass der konkrete Inhalt eines Textes um einen Themenkern „entfaltet“ wird. Setzen sich mehrere Texte in einer Textkollektion mit einem Themenkern auseinander, sei es auch in unterschiedlicher Entfaltung, so kann dieser Kern durch die angesprochene Reduktion und Konzentration auf Makropropositionen erkannt werden. Diese Sichtweise zeigt, dass es Zusammenhänge zwischen Texten gibt, die sich als Thema interpretieren lassen und durch das Auftreten von wiederkehrenden Strukturen und Wörtern manifestiert sind.

Mit der Theorie der funktionalen Satzperspektive findet sich eine weitere Sichtweise auf die thematische Analyse von Texten. Diese Theorie konzentriert sich auf zwei Funktionen eines Textes, die als „Thema“ und „Rhema“ bezeichnet werden (Lötscher (1987, vgl. S. 14), Danes u. Viehweger (1976, vgl. S. 29 ff.), Eroms (1986, vgl. S. 5)). Das Thema übernimmt die Funktion des dem Leser bekannten Konzepts und des Bezugspunktes (Eroms, 1986, vgl. S. 5), wohingegen das Rhema diesen Bezugspunkt ausführt und somit unbekanntes oder zu erklärendes zum Bezugspunkt beisteuert. Die Konzentration liegt bei diesem Ansatz auf der Satzperspektive. Die Darstellung des Textthemas ist immer eine Zerlegung aller Sätze in die Funktionen Thema und Rhema. Ein Bestandteil dieser Darstellung ist, dass durch die Zerlegung und Darstellung festgestellt werden kann, wie sich der Text thematisch entwickelt. Dies wird als thematische Progression (Eroms, 1986, vgl. S. 91) bezeichnet, welche in verschiedene Typen unterteilt wird.[4] In dieser Sichtweise ist ein Satz und sein Thema inklusive des Rhema eine mögliche Realisierung von vielen möglichen Realisierungen eines Themas. Die thematische Progression entspricht, in der Betrachtung des Themas als Themenkern, der „Entfaltung“ mit dem Unterschied, dass die thematische Progression möglichst vollständig abgebildet werden soll. Der starke Satzbezug und das Fehlen von Abstraktionsoperationen zeigt, dass diese Betrachtungsweise für Einzeltexte geeignet ist, jedoch nicht für die Generalisierung des Themas, um thematische Gemeinsamkeiten zwischen verschiedenen Texten darzustellen.

Eine dritte Gruppe linguistischer Betrachtungen auf das Textthema bilden die kategoriale Abgrenzung bzw. die Bezugstheorie und die Betrachtung des Themas als Fokus eines oder mehrerer Texte. Diese Herangehensweisen sind sehr ähnlich in ihrer Erfassung eines Textthemas. Sie unterschieden sich jedoch in der Darstellung und Interpretation. Die kategoriale Abgrenzung und die Bezugstheorie gehen davon aus, dass Textthemen aus sogenannten Nominalgruppen bestehen (Fritz (1982, vgl. S. 211), Lötscher (1987, vgl. S. 6)). Das bedeutet, dass sich Texte auf Objekte, die außerhalb der Texte existieren, beziehen und somit Referenzen auf diese Objekte bil-

[4] Es können grob drei Typen unterschieden werden (Eroms, 1986, vgl. S. 91).

- Lineare Progression: Das Rhema eines Satzes wird zum Thema eines folgenden Satzes.
- Durchlaufendes Thema: Das Thema wird über mehrere Sätze konstant beibehalten.
- Thematischer Sprung: Das Thema eines folgenden Satzes hat nichts mit dem Rhema oder Thema des bisherigen Textes zu tun.

In der Literatur werden zwar fünf Typen genannt von denen allerdings die verbleibenden zwei Varianten der hier dargestellten Typen sind.

den. Im Gegensatz zur Analyse der Propositionen wird nicht erfasst, wie die Objekte im Text verwendet werden. Da nach Fritz (1982) alle möglichen Beurteilungen und Entfaltungen in einem Text zugelassen sind, ist das Thema eben nicht die Erfassung aller semantischen Bezüge im Einzeltext, sondern nur die Extraktion der Referenzen. Damit wird die Darstellung der inneren Struktur eines Themas eingeschränkt (Lötscher, 1987, vgl. S. 14). Allerdings wird die Erfassung des Themas auf verschiedene Entfaltungen innerhalb vieler Texte erweitert. Es ist jedoch schwierig zu erkennen, welche Objekte, oder in diesem Fall Nomen, als Referenz im Text dienen und wie sie erfasst werden können. Vor allem ist unklar, ob ein Nomen stellvertretend ein Thema eines Textes bildet oder die Darstellung einer Nominalgruppe. Wird eine Gruppe von gleichwertigen Nomen dargestellt, so ergibt sich das Problem, dass die Relevanz einzelner Bezüge verwaschen wird und das Thema ungenügend erklärt wird. Um festzulegen, welche Objekte relevant für eine Erfassung sind, wird von Bayer (1980) vorgeschlagen, dass alle Objekte erfasst werden, die eine feste Denotation haben, also eine feststehende Bedeutung außerhalb des Textes, was wiederum Parallelen zum Begriff des Themas in der Thema-Rhema Gliederung zeigt. Das Problem der Erfassung mehrerer solcher Objekte im Text bleibt aber bestehen, da noch entschieden werden muss, wie die Wertigkeit dieser Objekte im Text beurteilt werden kann. Bayer (1980) schlägt vor, die Nomen nach einer Berechnung zu werten, um somit eine gewichtete Liste von Nomen für die Beurteilung des Textthemas zu erhalten. In dieser Berechnung der Nomengewichte spielen allerdings nicht die reinen Häufigkeiten der Nomen eine Rolle, sondern deren gegenseitige Bezüge und kontextuelle Einbettung. Relativ ähnlich zur Darstellung von Nominalgruppen, kann auch die Fokustheorie gesehen werden. Diese spiegelt die Thematik als Faktor wieder, der einen generellen Zusammenhang definiert. Dieser Faktor wird als Auswahlfunktion gesehen, welche die Thematik und die Ziele eines Textes enthält (Lötscher, 1987, vgl. S. 21 f.). Ob die Ziele des Textes beispielsweise Deskription oder Manipulation sind, verändert den zugrundeliegenden Faktor, was eine psychologische Komponente in diese Betrachtung mit einbringt. In der Kommunikationswissenschaft wird die Kombination aus Thematik und Handlungsabsicht als Framing bezeichnet und die Fokustheorie weist auf die Konzentration eines Textes auf Haupt- und Nebenthemen hin, wobei die Nebenthemen das Hauptthema gewissermaßen einrahmen.[5]

[5] Ein Beispiel hierfür wäre das Haupthema eines Textes, welches sich mit Automobilen beschäftigt. Je nachdem welche Aussage gemacht werden soll, wird das Hauptthema entweder mit Umweltaspekten wie dem Verbrauch und Energie ergänzt oder der Text erzeugt Wirkung mit Nebenthemen wie Leistung, Männlichkeit oder Preisen.

Hat der Leser es mit argumentativen Texten zu tun, so kann die Suche nach dem Thema nicht mit Propositionen oder Nominalgruppen beantwortet werden. In dieser Textsorte tauchen nämlich, durch den Einsatz verschiedener Argumente und Bezüge ganz unterschiedliche Entfaltungen und Rahmungen bzw. Faktoren auf, die nicht durch Mechanismen wie die thematische Progression zu fassen sind. In diesem Fall ist es sinnvoll die Themen in Form einer Fragestellung zu erfassen, sodass alle Meinungen oder Argumente zu einem Thema eliminiert werden (Lötscher, 1987, vgl. S. 25 f.).[6] Eine quantitative Beurteilung der Referenzen im Text spielt weniger eine Rolle, als die intellektuelle Beurteilung und Eingrenzung des zu verhandelnden Objekts, was wiederum sehr subjektiv ist.

Synthese linguistischer Themenanalysen

In einer oberflächlichen Betrachtung der dargestellten Theorien zur Themenanalyse, lassen sich Übereinstimmungen in der Art und Weise der Themeninterpretation feststellen. Dennoch existieren wesentliche Unterschiede innerhalb der vorgestellten Sichtweisen. Im folgenden Abschnitt sollen die Unterschiede und Gemeinsamkeiten noch einmal diskutiert werden, um darauf die Diskussion der technischen Verfahren aufzubauen, die in der Arbeit für die Themenanalyse vorgeschlagen werden sollen.

Einigkeit zwischen den einzelnen Theorien besteht durchaus, da alle zur „Konzentration und Abstraktion des gesamten Textinhalts“ (Mackeldey, 1987, S. 39) geeignet sind. Es gehen unterschiedliche Aspekte der Thematisierung verloren oder werden abstrahiert. Eine weitere Unterscheidung teilt die Eignung der Theorien in Analysen für mehrere Texte oder die ausschließliche Einzeltextanalyse. Eine generalisierte Definition, die auf Grundlage der genannten Theorien entstanden ist, hilft die Unterschiede deutlicher zu machen. „Das Thema eines Textes ist ein in irgendeiner Beziehung mangelhaftes Objekt, dessen Mangel in der Behandlung in diesem Text beseitigt werden soll.“ (Lötscher, 1987, S. 84) Die Einführung eines Mangels, der erklärt werden muss, erlaubt die Anwendung dieser Definition auf alle Theorien. Mit dieser Generalisierung ergibt sich jedoch eine Einschränkung: „Die inhaltliche Angabe eines Themas in einem Text kann also nicht nur in der Benennung des thematisierten Objekts allein bestehen, sondern muss auch in der Beschreibung der Qualität seines Mangels und der Charakterisierung der Mängelbehebung bestehen. Die Form der rein inhaltlichen Charakterisierung eines Themas ist der Charakterisierung der textgrammatischen

[6] Bei einer Diskussion über den Preis eines Produkts P bei der ein Teilnehmer die Position „günstig“ einnimmt und ein anderer die Position „teuer“, kann das Thema als Frage „Wie ist der Preis des Produkts P?“ ausgedrückt werden.

Funktion des Themas untergeordnet.“ (Lötscher, 1987, S. 108). Diese Festlegung schließt die Referenz- und Bezugstheorie der kategorialen Abgrenzung nicht mit ein, da hier Themen nur unter Angabe von Nominalgruppen dargestellt werden. Auch bei der Formulierung des Textthemas als Frage oder bei der Fokustheorie konzentriert sich die Analyse auf die Haupt- und Nebenobjekte in einem Text. Die eigentliche Ausführung der Thematik geht verloren. Im Gegensatz dazu behalten funktionale Satzperspektive und die Themenkerntheorie die Entfaltung eines Themas immer im Blickpunkt. Die Darstellung dieser Themenbetrachtungen bezieht sich aber meist auf Einzeltextanalysen. Die Erweiterung auf die Analyse mehrerer Texte oder kompletter Korpora ist bei den zuletzt genannten Theorien teilweise möglich, funktioniert aber nicht problemlos. Die Extraktion von Nominalgruppen aus mehreren Einzeltexten ist dagegen problemlos möglich. Unklar bleibt jedoch, wie die extrahierten Gruppen aus mehreren Texten zusammengefasst werden können, um eine globale Themenstruktur daraus zu bilden. Auch die Themenkerntheorie, die über Makrostrukturen bzw. Propositionen abstrahiert, ist von diesem Problem betroffen. Es ist zwar vorstellbar, dass diese Abstraktionen über mehrere Texte getroffen werden können. Dennoch müssen für die Beurteilungen thematisch unterschiedlicher Texte mehrere Abstraktionsgruppen gefunden werden, wobei hier theoretisch nicht ersichtlich wird, wie dies zu vollziehen wäre. In gleicher Weise ist vorstellbar, die weniger formal ausgedrückten Theorien über den Themenfokus oder die Darstellung des Themas als Fragestellung auf mehrere Texte zu erweitern. Einzig die Erweiterung der funktionellen Satzperspektive gestaltet sich durch die detaillierte Darstellung des Textes als thematische Progression schwierig. Bei dieser Art der Betrachtung wird keinerlei Abstraktion vorgeschlagen. Die individuelle Ausführung eines Textes mit anderen in Verbindung zu bringen, ist damit kaum möglich.

In der Betrachtung der thematischen Strukturen stellt Firbas (1992, S. 7) fest, „the elements of a clause, independent or dependent, differ in the extent to which they contibute towards the further development of the communication“. Nach dieser Feststellung sind Teile der Kommunikation mehr oder weniger dynamisch. Dieses Phänomen wird als kommunikative Dynamik bezeichnet. Eine Einsicht aus der Analyse der kommunikativen Dynamik ist die grundsätzliche Unterscheidung zwischen einer hohen und niedrigen Dynamik als Ausdruck einer relationalen Unterscheidung, die nicht auf absoluten Messzahlen beruht. Gleichzeitig wird aber der Unterschied zwischen einer hohen und niedrigen Dynamik gleichgesetzt mit dem Gegensatz zwischen alt und neu oder bekannt und unbekannt (Firbas, 1992, vgl. S. 106). Da in der Theorie der funktionalen Satzperspektive die Gliederung in Thema und Rhema

in den Gegensatz zwischen bekannt und unbekannt eingebettet werden kann, ergibt sich daraus eine interessante Ableitung. Wenn es Elemente in einem Text gibt, die sehr dynamisch sind und neues zu einem Thema beisteuern, so gibt es Elemente zu denen etwas Neues gesagt werden muss. Wenn mehrere Texte, wie zum Beispiel in der täglichen Berichterstattung, ein Ereignis oder ein Konzept besprechen, kann dieses Konzept jeweils unterschiedlich besprochen werden. Dennoch muss die Ansprache des Ereignisses oder Konzepts immer erfolgen, wenn der Text als Entfaltung einer Thematik verstanden werden soll. Die Einheiten im Text, die keiner hohen Dynamik unterliegen, müssen demnach in allen Texten, die ein Thema besprechen und auch so verstanden werden sollen, vorhanden sein. Somit ist vorauszusetzen, dass zumindest das Thema oder die Referenz auf das thematisierte Konzept in einer Textsammlung, in der es mehrere Texte zu einem Thema gibt, wiederholt wird.

Die Betrachtung der Theorien im Hinblick auf die Auswertung vieler Texte zeigt, dass durch vielfache Verwendung eines Themas in einer Textsammlung die Möglichkeit besteht, Themen mit Hilfe einer geringeren Dynamik bei der Referenzierung aus dem Text zu destillieren. Das bedeutet, dass es Wörter oder Konzepte gibt, die in mehren Texten angesprochen werden. Im Gegensatz dazu bilden Wörter und Konzepte, die nur in einzelnen Texten vorkommen, neue oder erklärende Inhalte ab. Konzepte, die in mehreren Texten verwendet werden, können ein gemeinsames Thema darstellen. Je größer die Anzahl der Texte ist, denen ein Thema zugesprochen wird, desto vielfältiger werden die Entfaltungen und damit der Anteil der neuen und dynamischen Informationen zu einem Thema erwartet. Dies bedeutet für die Themenanalyse von großen Textkollektionen, dass die spezifische Erläuterung eines Themas, wenn es in vielen Texten erfasst und zusammengefasst werden soll, immer mehr verwäscht und letztlich Abstraktionen gebildet werden, die die detaillierte Interpretation nahe des Textes nicht mehr erlauben und im Wesentlichen aus den wenig dynamischen Konzepten innerhalb der Texte bestehen. Diese Annahme weist neben der Themenkerntheorie darauf hin, dass es für die Darstellung von Themen verschiedene Abstraktionsgrade gibt, die immer genereller werden, bis nur noch das Konzept eines Themas erkennbar ist und nicht mehr der Kontext. Das Thema wird mit dem Abstraktionsgrad immer offener. Im Detail muss jedoch auf die einzelnen Entfaltungen zugegriffen werden, um die Diversität eines Themas erfassen zu können. Deshalb muss bei der Themenanalyse das Konzept des close- und des distant-reading nach Moretti (2005) angewendet werden.

Am Anfang der hier geführten Diskussion wird definiert, dass Themen immer im Zusammenhang mit Ereignissen, gesellschaftlichen Diskursen und gesellschaftli-

chen Feldern steht. Abschließend wird gesagt, dass die Abstraktionsniveaus einer Themenbetrachtung auf mehreren Texten genau diese Trennung erlauben, indem beispielsweise Abstraktionen für Einzelereignisse oder ganze Sachbereiche definiert werden. Je mehr Details einer Thematisierung ausgeblendet werden, desto genereller wirkt die thematische Kategorie. Dies lässt sich am besten anhand der Strukturierung von Nachrichtenartikeln in Zeitungen zeigen. Angefangen von Einzeltexten über Ereignisse, die wiederum in Rubriken und Ressorts eingeteilt werden, werden Auseinandersetzungen und Kommentare veröffentlicht, die eine gesellschaftliche Diskussion jenseits der zugrunde liegenden Ereignisse erläutern. Eine Story zieht demnach eine vielfältige Diskussion mit sich, die in gröberen Kategorien, wie zum Beispiel die Ressorts (gesellschaftliche Bereiche) Politik, Sport oder Kultur eingebettet sind.

Das Thema im zeitlichen Verlauf

Die Bestimmung der Themenstruktur innerhalb diachroner Textkollektion bildet die Grundlage von Themenanalysen in der Kommunikationsforschung. Diachron bedeutet, dass die Texte innerhalb verschiedener Zeitpunkte verfasst werden und der Verlauf von Bedeutungen, Relevanz und Spracheigenschaften der Texte untersucht werden kann (Saussure, 2001, vgl. S. 167 ff.). Durch die Kodierung der Themen in den Einzeltexten ist es möglich, Stichproben für unterschiedliche Messzeitpunkte zu bestimmen, wenn die Erscheinungsdaten der Texte zur Verfügung stehen (Schenk, 2007, vgl. S. 457). Dies ermöglicht neben der Analyse der Themenstruktur auch die Analyse von Trends in den Kollektionen. Merkmale eines Objekts, beispielsweise die Berichterstattungsmenge eines Themas zu verschiedenen Zeitpunkten, werden miteinander verglichen (Merten, 1995, vgl. S. 150). In Neuendorf (2002, vgl. S. 176) sind Beispiele für Querschnittsmessungen innerhalb von Inhaltsanalysen aufgeführt. Anhand der Beispiele wird gezeigt, welche Aussagen mit zeitabhängigen Analysen gemacht werden können. Danach können

- Messungen zeitversetzter Korrelationen und Beziehungen zwischen Themen,
- Messungen von Beziehungen zwischen einer Berichterstattung und öffentlicher Meinung (Agenda Setting) und
- die Bestimmung von Zusammenhängen zwischen der Berichterstattung und dem Sozialverhalten und -gewohnheiten durchgeführt werden.

Für zeitabhängige Betrachtungen kann eine retrospektive Analyseform gewählt werden. Für andere Fragestellungen können aber auch begleitende prospektive Analysen

durchgeführt werden. Diese werden unter der Bezeichnung Monitoring oder Clipping geführt (Bonfadelli, 2002, vgl. S. 53, vgl. S. 181). Begleitende Analysen werden angefertigt, um entstehende Trends und aktuelle Kennzahlen in veränderlichen diachronen Textkollektionen zu erkennen. Wie bei retrospektiven Messungen, werden Merkmale eines Objekts fortlaufend neu bestimmt. Die aktuellen Merkmale werden mit vergangenen Messungen verglichen. Die gemessenen Korrelationen oder Veränderungen, beziehen sich immer auf aktuelle Zeitpunkte. Retrospektive Analysen werden vor allem genutzt, um Entwicklungen und Phänomene nachträglich zu verstehen. Die Medienresonanzanalyse ist wichtige Anwendung prospektiver Studien. Sie wird verwendet, um beispielsweise die Reaktion der Medien auf eine lancierte Kommunikation oder Produkteinführung zu messen und zu evaluieren.

Nachrichtenfaktoren

Innerhalb von Themen können, wie bereits erwähnt, verschiedene Merkmale bestimmt werden, die durch Messungen zu unterschiedlichen Zeitpunkten in Längsschnitte übertragen werden können. Innerhalb von Themenstrukturen, die beispielsweise auf Ereignissen basieren können, gibt es Nachrichtenfaktoren, die Relevanz für den Wert einer Nachricht haben. Die Nachrichtenfaktoren können als zentrale Merkmale von Themen oder Nachrichten angesehen werden, da nach diesen Faktoren Ereignisse und Themen für die Berichterstattung ausgewählt werden. Durch die Bündelung mehrerer Faktoren innerhalb einer Nachricht, erhöht sich die Chance für eine journalistische Aufnahme (Schenk, 2007, vgl. S. 444). Um das öffentliche oder das journalistische Interesse an einem Thema zu erklären oder den Wert eines Themas zu bestimmen, ist die Analyse von Nachrichtenfaktoren zentral. Grundlage dieser Analyse sind aus den Nachrichten abgeleitete Regeln, nach denen Nachrichten ausgewählt werden (Rössler (2005, vgl. S. 227 ff.),Kepplinger (2011)). Die Nachrichtenfaktoren werden aufgeteilt in (Schenk, 2007, vgl. S. 444)

- die Dauer der Thematisierung (Zeit),
- die räumliche, politische und kulturelle Nähe,
- den regionalen, nationalen oder persönlichen Einfluss der Handlungsträger (Status),
- die Intensität und die Überraschung (Dynamik),
- das Konflikt-, Schadens-, und Erfolgspotential (Valenz) und

- die Identifikation und Personalisierung.

Viele dieser Merkmale lassen sich durch quantitative Methoden bestimmen. In Abschnitt 2.2 wird näher erläutert, welche Möglichkeiten automatisierte Methoden bieten, um die Bestimmung von Nachrichtenfaktoren in großen Textkollektionen zu unterstützen. An dieser Stelle sollen nur die Themenintensität und die Handlungsträger angesprochen werden, um die Bedeutung der Längsschnittmessungen für die Beurteilung von Themen zu verdeutlichen.

Die Intensität eines Themas wird unter anderem bestimmt, indem innerhalb der Stichprobe ausgezählt wird, welche Artikel das Thema oder eine assoziierte Kategorie enthalten. Innerhalb aller Artikel oder Kodiereinheiten, denen ein Thema zugeordnet wurde, kann bestimmt werden, welche Personen oder Akteure auftreten, sodass eine Zuordnung der Handlungsträger möglich ist. Die Messung der Intensität und der Handlungsträger kann in Stichproben zu unterschiedlichen Zeitpunkten durchgeführt werden und so in einen Längsschnitt übertragen werden. Über die so generierten Zeitreihen lassen sich thematische Profile bestimmen, welche den Verlauf der Themen in Bezug auf unterschiedliche Merkmale darstellen.

Die Dynamik, der die Merkmale einer Thematisierung über die Zeit unterliegen, zeigt, dass Themen Einflussfaktoren unterworfen sind. Eine Thematisierung wird durch Ereignisse wichtig oder, wenn es keine neuen Informationen gibt, durch andere Ereignisse und daraus folgenden Themen ersetzt. Aus der Beobachtung der Themendynamik wird das Verhalten von Themen erklärt und es können Phasen der Thematisierung abgeleitet werden (Kolb, 2005). Ausgehend von der Idee, dass Themen um Aufmerksamkeit ringen, wurde ein Themenzyklus entwickelt, der dem Produktlebenszyklus aus der Ökonomie nachempfunden ist (Fischer, 2001; Höft, 1992). Nach diesem Modell unterscheidet Kolb (2005, vgl. S. 80 f., vgl. S. 92 f) folgende Phasen.

1. **Latenzphase:** In dieser Phase tritt das Thema erstmals auf und das Interesse ist nicht hoch. Das Thema ist interessant für Experten und wird von einem wissenschaftlichen Austausch dominiert(Kolb, 2005, vgl. S. 80).

2. **Schlüsselereignisse:** Durch verschiedene Ereignisse gewinnt das Thema an Relevanz für gesellschaftliche Bereiche und die Medien berichten stärker darüber.

3. **Aufschwung:** Durch die gesellschaftliche Relevanz werden die themenbezogenen Probleme zu gesellschaftlichen Problemen und der Wert des Themas für die Medien steigt weiter an.

4. **Etablierung, Politisierung:** Die Medien berichten über ein populäres Thema, während sich politische Akteure an der Thematisierung beteiligen und das Thema weiter gestalten. Die Akteure ringen in dieser Phase um die besseren Argumente.

5. **Entscheidung und Lösung:** Durch die komplette Erklärung und Einführung des Themas und dessen eigenen Problemen und Akteuren kommt es zu Entscheidungen, wie z.B. der Verabschiedung von Gesetzen (Kolb, 2005, vgl. S. 227), die Probleme des Themas lösen und Diskussionen beenden.

6. **Abschwung, Marginalisierung:** Findet ein Thema eine Lösung, so geht die Aktivität der Akteure und das Interesse der Medien an einem Thema zurück, sodass die Thematisierung dauerhaft verschwindet.

An der Themenintensität kann abgelesen werden, wann ein Thema an Wert für die öffentliche Berichterstattung gewinnt. Abhängig ist dies meist von Ereignissen oder Entscheidungen, die innerhalb einer Thematisierung kommuniziert werden. Die Übergänge zwischen den Phasen oder auch eine Progression der Thematisierung hängt so von Schlüsselereignissen ab. Für abgeschlossene Thematisierungen gibt es nach der Phaseneinteilung einen idealtypischen Verlauf für Themen. Dieser Themenzyklus kann für die strukturierte Themenanalyse verwendet werden. Die Anwendung des Phasenmodells führt jedoch zu Problemen, da der idealtypische Verlauf nicht auf alle Thematisierungen passt. So wurde das Modell beispielsweise für den Spezialfall der Kriegsberichterstattung angepasst (Miltner u. Waldherr, 2013).

Die Anwendung des Themenzyklus lässt zu, dass durch die Beobachtung von Merkmalen eines Themas auf einzelne Phasen geschlossen werden kann. Der Verlauf wird nach Kolb sehr idealtypisch dargestellt und passt nicht auf alle Thematisierungen. Der Zyklus kann unterschiedlich sein und so kommt es bei der Anwendung nicht auf die exakte Herstellung des idealen Themenzyklus an, wie in Abbildung 2.1 dargestellt. Zentral ist die Möglichkeit, die Phasen eines Themenzyklus zu identifizieren und die Eigenschaften der Phasen zu nutzen, um eine untersuchte Thematisierungen zu verstehen. Die Arbeit mit den Zyklen schafft die Möglichkeit, Themen und deren Verlauf vergleichbar zu machen. Prognosen können erstellt werden oder Phasen anderer Themen verglichen werden. Durch die Identifikation der Phasen können zielgerichtet Nachrichtenwerte, Frames (Rössler, 2005, vgl. S. 230) oder Agenda-Setting Effekte ausgewertet werden. Somit lässt sich durch die Phasen nicht nur die Aufmerksamkeit, welche ein Thema in den Medien erfährt, darstellen. Vielmehr können in Dokumenten, die zu einer bestimmten Phase gehören, temporäre Aspekte eines Themas bestimmt

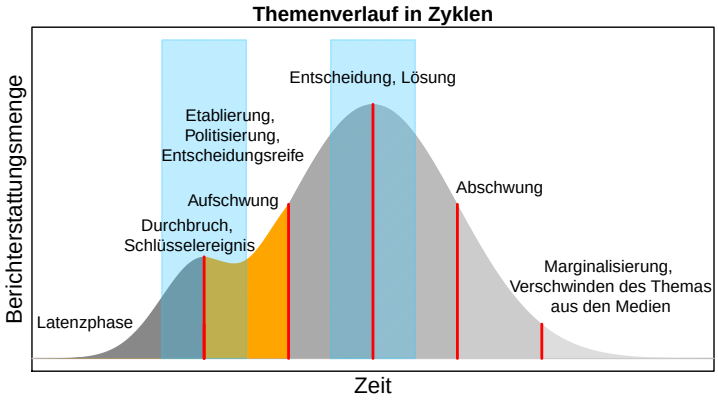

Abbildung 2.1: Thematisierung in Zyklen nach (Kolb, 2005) und Skizze einer möglichen Auswahl von Dokumenten und Inhalten aufgrund von Schlüsselereignissen, die zu einer Phase im Themenzyklus gehören.

werden. Nach diesem Modell lässt sich ein Thema als sequenzielle Themengeschichte darstellen. Für unternehmerische Kommunikation kann es große Vorteile bringen, wenn sich die besten Zeitpunkte für eine lancierte Kommunikation oder für den Eintritt bestimmter Akteure in die Kommunikation aus dem Themenverlauf ablesen lassen. Auch die Früherkennung von Trends und die Kenntnis über mögliche Auswirkungen anhand der Zyklenbetrachtung kann für die Unternehmenskommunikation von Vorteil sein.

Um die Möglichkeiten zur Themeninterpretation mit Hilfe der Zyklenbetrachtung zu nutzen, müssen Merkmale aus Textkollektionen extrahiert werden, die es erlauben, die Phasen abzubilden. In dieser Arbeit soll ein Ansatz verfolgt werden, Themenzyklen und deren Eigenschaften durch Methoden der automatischen Sprachverarbeitung und des Text-Mining aus Textkollektionen zu extrahieren. Es soll untersucht werden, wie Aggregation der Themenstrukturen innerhalb verschiedener Phasen zu bilden sind, sodass die Themenanalyse in großen Textmengen unterstützt werden kann. Durch die automatische Themenzuordnung von Dokumenten in großen diachronen Korpora, sollen Längsschnitte für die thematischen Anteile an der Berichterstattungsmenge gebildet werden, um so Themenzyklen sichtbar zu machen. Durch die Identifikation der Themen und ihrer Phasen sollen Entwicklungen in den Merkmalen eines Themas erkennbar werden. In Abbildung 2.1 wird skizziert, wie Phasen einer Thematisierung genutzt werden, um z.B. Dokumente, Akteure oder Wortverwendungen zu extrahieren, die einer bestimmten Phase des Themenzyklus zugeordnet werden. Mit der Kombination aus automatischen Methoden und dem theoretischen Ansatz der Themenzyklen kann so ein neuer Ansatz entwickelt werden, Dokumente aus einer

diachronen Textkollektion mit dem Fokus auf eine bestimmte Phase einer Thematisierung zu extrahieren.

Ökonomische Probleme der Inhaltsanalyse

Inhaltsanalysen in großen und wachsenden Textkollektionen stellen mit den vorgestellten Methoden der Inhaltsanalyse eine große Herausforderung dar. Da die Kodierung der Kategorien oder Themen meist manuell vorgenommen wird, stellen große Textmengen ein Problem dar. Zum einen kann nicht genau gesagt werden, welche Texte überhaupt relevant für eine Fragestellung sind. Zum anderen ist die Kodierung der relevanten Texte, durch deren Vielzahl, kaum zu bewältigen. Mit herkömmlichen Methoden muss sich die Analyse demnach auf einen Ausschnitt der Grundgesamtheit stützen, was insbesondere bei der Entwicklung der Kategorien problematisch sein kann, da nicht alle Ausprägungen einer Kategorie vorliegen. Die Art und Weise, wie die Zufallsstichprobe entsteht, beeinflusst die Ergebnisse.

Die Kodierung der Texte wird meist durch mehrere Mitarbeiter vollzogen und dauert durch den manuellen Prozess lange. Je mehr Text bearbeitet werden muss, desto länger können sich Projekte ziehen. Dauer und Kosten sind in der Inhaltsanalyse durchaus projektrelevant und beeinflussen so die zu erreichenden Ergebnisse. Das Anliegen dieser Arbeit soll im Hinblick auf die angesprochenen Probleme auch ein Beitrag sein, diese Probleme für Themenanalysen zu lösen und diese effizienter für den Einsatz in großen Textkollektionen zu machen.

2.2 Computergestützte Analyse digitaler Textquellen

Um die inhaltsanalytische Aufgabe der Themenanalyse in digitalen Textkollektionen automatisch durchzuführen, muss der Analyseprozess in der Lage sein, digitalisierte Textdokumente zu verarbeiten. Grundsätzliche Bestandteile von Texten, wie Wörter oder Sätze, müssen erkannt werden und für eine weitere Bearbeitung im Rechner verfügbar sein. Aspekte der inhaltlichen Verarbeitung, der Datenstruktur und der Datenhaltung spielen eine Rolle. Diese Aufgaben werden in einem Teilgebiet der angewandten Informatik, der automatischen Sprachverarbeitung, erforscht. Die automatische Sprachverarbeitung bietet insbesondere Methoden für die Verarbeitung digitaler Textdokumente an. Der folgende Abschnitt bezieht sich deshalb auf die Methoden der automatischen Sprachverarbeitung, die sich mit der kognitiven Seite des menschlichen Sprachverständnis, mit dem Verständnis der Beziehungen linguistischer Äußerungen und dem Verständnis linguistischer Strukturen beschäftigen (Manning,

1999, vgl. S. 3). Um Textdokumente verarbeiten zu können, müssen diese in geeigneter Form verarbeitet und gespeichert werden (Heyer, 2006, vgl. S. 51). Deshalb werden im Folgenden grundsätzliche Datenhaltungskonzepte bei der Verarbeitung von Textkollektionen angesprochen.

2.2.1 Verarbeitung und Repräsentation

Damit eine Textkollektion Informationen für die Inhaltsanalyse bereitstellen kann, müssen diese Informationen aus den meist unstrukturieren und nicht annotierten digitalen Textquellen erzeugt werden. Dazu gehört die Identifikation von Wörtern, das Festlegen von Satzgrenzen, die Erkennung von Kapiteln und Abschnitten (Heyer, 2006, vgl. S. 57) und die Verknüpfung von Metadaten.

Quellen, Zeichensätze und Sprachen

Die Ausgangslage einer jeden Textkollektion bilden die Quelldokumente, aus denen der Text extrahiert wird. Diese Quelldokumente können strukturierte Datensätze oder unstrukturierte Texte sein.[7] Je weniger Struktur eine Textquelle aufweist, desto schwieriger ist der Zugriff auf den darin enthaltenen relevanten Text. Selbst mit Hilfe einer Struktur ist es oft schwierig, den darin enthaltenen relevanten Text zu selektieren. So ist beispielsweise die Extraktion der relevanten Textteile einer HTML-Seite schwierig, da oft Werbung oder Strukturinformationen in einem Dokument integriert sind. Um dennoch an die Texte zu kommen und nicht relevante Inhalte auszublenden, existieren verschiedene Ansätze. Ein Ansatz besteht darin, mögliche Muster von nicht relevantem Text zu finden und über reguläre Ausdrücke zu selektieren. Damit ist es möglich, Inhalte wie Navigationsstrukturen oder Werbung zu identifizieren. Weiterhin können wiederholte Inhalte wie Kopfzeilen, Fußzeilen und Seitenzahlen über reguläre Ausdrücke selektiert werden. Handelt es sich um strukturierte Dokumente, so kann die semantische Bedeutung der Auszeichnungen und Annotationen genutzt werden, um die Inhalte zu selektieren. Im Idealfall ist die Textkollektion bereits erstellt oder die Quellen liegen in einer bekannten Struktur vor, sodass die Identifikation der Nutztexte und der Metadaten problemlos ist. Sind die zu benutzenden Quellen gänzlich unstrukturiert, müssen relevante Textteile in eine Struktur überführt und unwichtige Textteile ignoriert werden. Bei gescannten Dokumenten, die mit einer OCR-Erkennung in Textdokumente umgewandelt werden, kann es außerdem vorkom-

[7] Strukturierte Datensätze können Dokumente sein, die eine strukturierte Auszeichnung wie XML nutzen oder einzelne Texteinheiten im Text expliziert auszeichnen.

men, dass jeder Umbruch der gedruckten Zeilen in der digitalen Datei einkodiert ist. Hier muss versucht werden, die unnötigen Umbrüche zu entfernen und nur Umbrüche zu erhalten oder zu nutzen, die ausschließlich Informationen über Absätze oder Überschriften wiedergeben.

Wie bereits angesprochen, teilen sich die Arten der Textkollektionen in strukturierte und unstrukturierte Quellen auf. Für unterschiedliche Aufgaben werden verschiedene Quellen benötigt, die an dieser Stelle systematisch dargestellt werden sollen.

- **Internet:** Die im Internet befindlichen Informationen sind für die Analyse aktueller Entwicklungen und Trends besonders wertvoll, da hier zeitnah Informationen bereit gestellt werden. Die öffentliche Berichterstattung spiegelt sich in den Online-Angeboten verschiedener Nachrichtenmedien wieder. Weitere Quellen, wie öffentlich zugängliche Foren, digitale Enzyklopädien, geschäftlich oder privat betriebene Informationsseiten ermöglichen einen Zugang zu speziellen und fachlichen Inhalten. Im Bereich der Nachrichtenmedien werden die öffentlichen Archive für die Anwendung in der Inhaltsanalyse immer wichtiger, da diese den kontrollierten Abruf der Nachrichtenquellen ermöglichen und die Artikel ständig zur Verfügung stehen. Nicht immer ist der Zugriff frei, aber in diesen Quellen kann strukturiert recherchiert und die Inhalte abgerufen werden. Manche der Archive sind über einen Internet-Auftritt organisiert und es ist mit einem gewissen Aufwand verbunden, die Textinhalte aus der Struktur des Auftritts zu selektieren. Wesentlich einfacher und zuverlässiger kann auf die Berichterstattung zugegriffen werden, wenn die Anbieter Web-Services nutzen, die die Inhalte in strukturierter Form exportieren und die Selektion von Metadaten und Text durch diese Strukturierung problemlos ist.[8]

- **Aufbereitete Korpora:** Verschiedene Organisationen bieten bereits aufbereitete Textsammlungen an (Heyer, 2006, vgl. S. 53), die bereits in einer Datenstruktur vorliegen. Solche abgeschlossenen und bereits verarbeiteten Textquellen werden Korpus genannt, eine Terminologie, die im Folgenden weiter verwendet wird. Es werden unterschiedliche Ressourcen für verschiedene Sprachen angeboten, die aus unterschiedlichen Quellen entnommen sind. Dies können aufbereitete Komplettausgaben einer Zeitung sein. Die Besonderheit dieser

[8] Beispiele für Nachrichtenarchive im Internet sind http://www.spiegel.de/nachrichtenarchiv/, http://www.zeit.de, http://open-platform.theguardian.com/ (Webservice) oder http://developer.nytimes.com/ (Webservice).

Textquellen ist, dass in den einzelnen Dokumenten keine unbrauchbaren Daten enthalten sind und der Nutztext direkt verarbeitet werden kann. Meist existieren zu allen Dokumenten vorbereitete Metadaten und linguistische Informationen. Diese Textsammlungen eignen sich besonders für retrospektive Betrachtungen, die nicht notwendigerweise tagesaktuelle Informationen enthalten müssen.

- **Lose Dokumentsammlungen:** Eine weitere Quelle für Analysen bieten lose Dokumentkollektionen. Diese bestehen aus einzelnen Datensätzen, die zusammen eine Textsammlung bilden. Meist enthalten Dokumente in solchen Sammlungen wenige strukturierte Informationen oder Metadaten und liegen meist als Text-, PDF oder Textverarbeitungsdateien vor.

Es existieren weltweit mehrere Standards, Zeichensätze mit Hilfe von Computern darzustellen (Stock, 2007, vgl. S. 117). So können die Quellen in unterschiedlichen Zeichensatzformaten wie beispielsweise ISO 8859-1, UTF-8 oder ANSI vorliegen. Bei der Verarbeitung von Textdokumenten muss darauf geachtet werden, dass alle genutzten Dokumente bei der Eingabe in einen einheitlichen Zeichensatz umgewandelt werden, ohne dass Zeichen verloren gehen oder nicht mehr lesbar sind. Weiterhin müssen alle Textdaten getrennt nach der verwendeten Sprache gespeichert werden oder eine Information über die Sprache enthalten. Viele statistische Methoden für die automatische Sprachverarbeitung können nicht mit unterschiedlichen Sprachen arbeiten und eine sprachliche Trennung oder eine Übertragung in eine Referenzsprache muss vor deren Anwendung erfolgen. Diese Art der Sprachunabhängigkeit soll im weiteren Verlauf aber nicht näher vertieft werden.

Vorbereitung der Texte

Liegt in einer Textkollektion nach der Verarbeitung der Eingabequellen der reine Nutztext vor, müssen die Bestandteile der Texte für eine computerlinguistische Analyse getrennt werden. Für inhaltsanalytische Aufgaben müssen die Texte mindestens in Einheiten zerlegt werden, die als Analyse- und Kodiereinheit definiert sind. In einer Inhaltsanalyse dienen die Einheiten Dokument, Absatz, Satz und Einzelwörter als Analyseeinheiten (Merten, 1995, vgl. S. 281). Wenn Phrasen als Einheit verwendet werden sollen, so müssen diese erst in ihrer Gestalt definiert werden. Dies ist meist erst nach einer Sichtung der Texte möglich und wird deshalb in diesem Abschnitt über die automatische Vorbereitung der Texte nicht weiter diskutiert.

In Abschnitt 2.1.3 wird deutlich, dass ein Thema durch die Zugehörigkeit einzelner Wörter, als Proposition, Nomen oder Faktor, bestimmt wird. Aus diesem Grund muss

eine Tokenisierung der Texte erfolgen, sodass die Wortgrenzen zur Bestimmung der Einzelwörter, die Tokens, vorliegen (Manning, 1999, vgl. S. 124). Meist wird die Regel eingesetzt, dass Wörter mit einem Leerzeichen getrennt sind oder bestimmte Sonderzeichen als Trenner interpretiert werden. Von Sprache zu Sprache gibt es allerdings Probleme und Sonderfälle. So werden Abkürzungen oft mit einem Punkt beendet oder Apostrophe eingesetzt. Da dies von Sprache zu Sprache unterschiedlich ist oder völlig andere Regeln gelten, wie in asiatischen Sprachen, muss die Tokenisierung für jede Sprache optimiert werden. In dieser Arbeit liegt die Konzentration auf der Analyse westeuropäischer Sprachen, welche sich mit der Regel, dass Wörter an einem Leerzeichen getrennt werden, gut tokenisieren lassen. Ein Ansatz, die Sonderregeln jeder Sprache zu lernen, ist das Trainieren eines Maximum Entropy Modells einer Sprache, sodass bei unsicheren Fällen ein Klassifikator über die Trennung der Wörter entscheidet. Dieses Verfahren ist als Software mit verschiedenen Tokenisierungsmodellen für unterschiedliche Sprachen implementiert (Reynar, 1998; openNLP).

Die Bestimmung der Satzgrenzen in einem Text kann nach einem ähnlichem Vorgehen wie die Bestimmung der Token erfolgen. Die Trennung orientiert sich an Satzzeichen einer Sprache, die in den Dokumenten verwendet wird (Manning, 1999, vgl. S. 134). Da Satzzeichen allerdings auch mehrdeutig verwendet werden, wie zum Beispiel bei Abkürzungen, müssen bei dieser Aufgabe Sonderregeln definiert werden. Einerseits können die Regeln fest vorgegeben werden (Manning, 1999, vgl. S. 135) oder es können Maximum Entropy Modelle mit den Regeln einer Sprache trainiert werden, um die Behandlung der Sonderfälle als Klassifikationsproblem zu lösen (Mikheev, 1998; openNLP).

Absätze stellen in Texten eine Möglichkeit dar, die Texte nach Aussagen und Sinneinheiten zu strukturieren. In digitalen Texten, die in strukturierter Form vorliegen, können die Absätze direkt aus der Struktur erkannt werden. Liegt eine Strukturierung nicht vor, so muss auf Muster zurückgegriffen werden oder die Absätze müssen über eine Phrasenerkennung identifiziert werden. Im Fall der Mustererkennung kann die Formatierung und einfache Strukturierung der Texte genutzt werden. Es kann beobachtet werden, dass die Absätze in unstrukturierten Texten oft über eine n-fache Verwendung von Zeilenumbrüchen getrennt sind. Besteht keine Möglichkeit sinnvolle Strukturen aus einem vorliegenden Textformat zu nutzen, so existieren Verfahren, die mit Algorithmen des maschinellen Lernens versuchen, Absatzgrenzen zu klassifizieren. Diese Ansätze nutzen Eigenschaften wie Satzentropien und Wortverteilungen, um über Boosting Algorithmen (Sporleder u. Lapata, 2004), Support Vector Machines

(SVM) (Fukumoto u. Suzuki, 2002) oder Perceptron Modelle (Genzel, 2005), um die Absätze automatisch zu bestimmen.

Während der Verarbeitung der Einzelwörter bzw. der Token in einem Text, kann die Wortform der Einzelwörter betrachtet werden. Oft sollen bei der Bildung von Statistiken nicht einzelne flektierte Wortformen analysiert werden, sondern Wörter eines Lemmas oder eines Wortstammes müssen gleich behandelt werden. Soll die Morphologie der Wörter in den Texten bei der Themenanalyse beachtet werden, so lohnt es sich, dies direkt bei der Vorverarbeitung zu tun. Beim Stemming wird versucht, Affixe der flektierten Wortformen abzuschneiden und auf einen Wortstamm zurückzuführen (Porter, 2001, 1997; Lovins, 1968; Hull, 1996). Bei der Lemmatisierung oder Grundformreduktion wird ein Wort auf den lexikalischen Ursprung zurückgeführt. Vollformenlexika mit flektierten Wortformen und deren Grundform werden für diese Aufgabe herangezogen (Kunze u. Lemnitzer, 2002; Fellbaum, 1998). Eine weitere wichtige Vorbereitung eines Textes stellt die Auszeichnung von Part-of-speech (POS) Informationen dar. Jedem Wort im Text wird eine Wortart automatisiert zugeordnet. Diese Informationen können im späteren Verlauf für die Analyse von Satzstrukturen, Wortartfilterung oder die Analyse funktionaler Eigenschaften der Wörter genutzt werden. POS Informationen werden meist, aufgrund von manuell erzeugten Trainingsdaten, automatisch klassifiziert und es existieren gute Werkzeuge für diese Aufgabe (openNLP). Weiterhin ist die „Named Entity Recognition“ ein wichtiger Bestandteil der Vorverarbeitung, da so Eigennamen für die weitere Verarbeitung zur Verfügung stehen. Dies ist für die Bildung von Wortgruppen wichtig, die Eigennamen repräsentieren, wie beispielsweise die gemeinsame Nennung von Vor- und Zunamen im Text. Diese Informationen sind wichtig, um Personengruppen oder Ortsbezüge innerhalb verschiedener Aufgaben mit nutzen zu können. Für diese Aufgabe existieren Ressourcen und Werkzeuge für die automatische Klassifikation (Finkel u. a., 2005; Faruqui u. Padó, 2010).

Bei der initialen Verarbeitung digitaler Texte müssen alle Informationen, die in einer automatisierten Themenanalyse erfasst werden sollen, in den Daten erkannt und in den Dokumenten annotiert werden. Das Verarbeiten und Selektieren definierter Informationen in Textdaten wird Informationsextraktion genannt (Manning, 1999, vgl. S. 376).

Speicherung verarbeiteter Texte und Metadaten

Die automatische Verarbeitung natürlichsprachiger Dokumente setzt voraus, dass die Quelldokumente in einer Vorverarbeitung vorbereitet werden. Damit ein späterer

Prozess auf die Vorverarbeitung zurück greifen kann, muss aus den Quelldokumenten in der Regel eine neue Datenbasis aufgebaut werden, die die Ergebnisse der Vorverarbeitung berücksichtigt. Soll eine große Anzahl von Textdokumenten für Textanalysen verfügbar gemacht werden, so muss die Art der Speicherung den späteren Zugriff nach verschiedenen Kriterien erlauben. Die Datenstruktur muss so beschaffen sein, dass

1. Metadaten der Artikel, wie Autorenschaft, Veröffentlichungsdatum, Publikation oder Seitenzahl, zur Verfügung stehen,
2. der Volltext der Dokumente gespeichert wird,
3. der Volltext so repräsentiert ist, dass der Zugriff auf einzelne Textelemente, wie Wörter, Sätze und Absätze, gegeben ist und
4. die Datenstruktur erlaubt, zusätzliche Informationen zum Text zu speichern, sodass Annotationen im Text vorgenommen werden können.

Grundsätzlich existieren verschiedene Strategien, die Daten vorzuhalten. Aus den unterschiedlichen Möglichkeiten muss unter Beachtung der Anforderungen einer Anwendung eine geeignete Vorgehensweise gewählt werden.

- **Tokenisierte und satzseparierte Textdateien:** Die einfachste Strategie, die Textdaten nach der Vorverarbeitung zu hinterlegen, ist die separierten Texteinheiten, wie Token, Sätze oder Absätze, innerhalb einer Textdatei abzutrennen. Die Token werden mit Leerzeichen getrennt und ein Absatz oder Satz auf je eine Zeile der Textdatei gesetzt. So ist die Trennung durch eine einfache Struktur jederzeit wieder lesbar. Ein Nachteil bei dieser Verarbeitung ist, dass zusätzliche Informationen über einzelne Token, wie beispielsweise POS-Tags nicht einfach eingebracht werden können. Für eine Zeile tokenisierten Text können eine Dokumentzugehörigkeit und entsprechende Metadaten hinterlegt werden. Die Verarbeitung solcher Formate ist einfach und benötigt keine zusätzlichen Technologien für die Verarbeitung.
- **Annotierter Text:** Um die Struktur eines Textes zu repräsentieren und um den originalen Text vollständig zu erhalten, werden die Strukturinformationen innerhalb der Dokumente annotiert. Für die Repräsentation der Volltexte, der Textstruktur und für Textannotationen bietet sich die Stand-Off Annotation an (Burghardt u. Wolff, 2009). Eine solche Struktur hinterlegt Informationen über einen Text nicht im Text selbst. Statt dessen wird ein externer Verweis auf eine

Spannweite im Fließtext gesetzt. Im Gegensatz zur Inline-Struktur (Heyer, 2006, vgl. S. 53) für Annotationen gibt es die Möglichkeit, den originalen Text zu erhalten und später weitere Informationen im Text speichern zu können. Über die Struktur ist der spätere Zugriff auf die Einheiten problemlos. Nachteilig ist der erweiterte Speicherplatz, der für die Annotationen benötigt wird, was bei großen Textkollektionen einen erhöhten Ressourcenbedarf voraussetzt.

- **Textdatenbanken und Volltextindex:** Wenn vorbereitete Dokumente als Textdatei oder strukturierte Datei vorgelegt werden, so ist der Zugriff auf die Einzeltexte oder die Selektion schwierig, da keine indizierte Datenstruktur für selektierbare Kriterien existiert. Die Lösung stellt die Übertragung der Daten in eine Textdatenbank dar, die über Indizes und eine relationale Struktur einen effektiven Zugriff auf die Texte gewährleisten kann (Heyer, 2006, vgl. S. 57 f.). Die extrahierten Einheiten, wie Wörter, Sätze und Absätze, können direkt in eine relationale Datenstruktur überführt werden. Über die Extraktion einer Wortliste (Heyer, 2006, vgl. S. 59) können inverse Listen (Baeza-Yates u. Ribeiro-Neto, 2011, vgl. S. 340 f.) in die relationale Struktur integriert werden. Diese werden genutzt, um Sätze, Absätze und Dokumente effizient nach Suchbegriffen zu selektieren. Meist werden bei der Erstellung der inversen Listen Textstatistiken, wie zum Beispiel Worthäufigkeiten, für die gesamte Textdatenbank extrahiert. Wenn eine Datenbank mit Textdokumenten in Stand-Off Annotation zur Verfügung steht, müssen diese in inverse Listen überführt werden, um eine Dokumentselektionen durchführen zu können. Dies erfordert allerdings, dass die relationale Struktur oder eine Indizierungsstruktur für die analytische Fragestellung und die Annotationen angelegt werden. Mitunter muss eine Struktur für jede Fragestellung oder Analyse neu definiert werden. Mit dieser Zugriffsstrategie lassen sich Volltextdatenbanken implementieren, die den gesamten Fließtext von Dokumenten enthalten und auf Suchbegriffen basierende Anfragen erlauben (Stock, 2007, vgl. S. 157).

Der statische Aufbau von Textdatenbanken kann die Vorteile der schnellen Suche oder Verdichtung von Informationen verlieren, wenn sich die Anforderungen an die Abfragen oft ändern. Dies ist bei Inhaltsanalysen der Fall, da jede Analyse mit neuen Hypothesen und Fragestellungen an die Texte herangeht. Mitunter müssen Statistiken aus Untermengen der Dokumente extrahiert werden. Die Statistiken müssen trotz einer vorhandenen Datenbank neu berechnet werden. Aus diesem Grund ist es notwendig, dass für inhaltsanalytische Aufgaben eine Datenstruktur gewählt wird, die schnell an neue Fragestellungen angepasst werden kann. Im Anhang A werden

Ansätze vorgestellt, wie die Textdaten für die Analysen vorgehalten und verarbeitet werden können. Die Grundüberlegungen, die für die Datenbasis angestellt werden müssen, werden aber hier schon kurz vorgestellt.

- **Abgeschlossenheit:** Für manche Aufgaben, wie das Clipping und Monitoring von Themen, ist es nötig, tagesaktuelle Dokumente vorzuhalten und zu analysieren. Das bedeutet, dass die Datenbasis mit der Zeit wächst und ältere Daten an Bedeutung für die Analyseaufgabe verlieren. Mit jeder neuen Dokumentmenge, die in das System eingebracht wird, ändern sich Textstatistiken und eine Datenhaltung muss in der Lage sein, auf diese Veränderungen zu reagieren oder diese Änderungen abzubilden.

- **Synchrone und diachrone Sprachwissenschaft:** Je nachdem, ob eine Textkollektion als diachron oder synchron betrachtet wird, wird eine andere Repräsentation der Daten benötigt. Denn jedes zu analysierende Glied in einer Abfolge von zeitlich abhängigen Textmengen muss in abgeschlossener Form untersucht werden, um Unterschiede herauszuarbeiten.

In Kapitel 4 werden Auszüge aus Zeitungskollektionen dargestellt. Die Textmenge einer solchen Quelle beträgt ca. 100 - 200 Dokumente pro Tag, die für eine Quelle verarbeitet werden müssen. Soll die Textquelle immer wieder durch neue Dokumente ergänzt werden, so muss diese Menge täglich verarbeitet und ergänzt werden können, wenn nicht nur repräsentative Auszüge bewertet werden sollen. Gewissermaßen ist die Vorverarbeitung und die Datenhaltung an dieser Stelle das Bindeglied zwischen möglichen Analysen und inhaltsanalytischen Anforderungen und ist neben der eigentlichen Analyse eine wichtige Arbeitsgrundlage.

2.2.2 Maschinelles Lernen (Machine-Learning) und Text-Mining

Die Verfahren und Modelle des maschinellen Lernens sind von besonderer Bedeutung für die Anwendung automatisierter Verfahren für die Inhaltsanalyse von Themenstrukturen. Diese Verfahren untersuchen Datenpopulationen automatisch nach erlernbaren Regeln. Dagegen erfordert die manuelle Erstellung solcher Regeln, ähnlich der Erstellung eines Codebuchs, einen hohen Aufwand, um die Komplexität aller möglichen Ausprägungen zu erfassen. Sind Regeln unvollständig, so werden falsche oder keine Inhalte erfasst. Die Verfahren des maschinellen Lernens können schneller an Veränderungen in den Datenquellen angepasst werden, sodass aufwändige Änderungen von Regeln keine Rolle spielen.

Bei der maschinellen Verarbeitung von Texten spielt der Einsatz des maschinellen Lernens (Machine-Learning) eine entscheidende Rolle. Unter dem Begriff werden informationstheoretische, statistische und mathematische Methoden zusammengefasst, mit denen aus vorhandenen Daten Zusammenhänge gelernt werden können (Hastie u. a., 2001, vgl. S. 1). Hinter der Anwendung und Entwicklung der Verfahren des maschinellen Lernens steht die Annahme, „dass es einen Prozess gibt, der die von uns beobachteten Daten erklärt." (Alpaydin, 2008, vgl. S. 1). Die Prozesse, auch als Modelle bezeichnet, die für das Lernen aus Daten entwickelt werden, sollen einerseits dazu dienen, Vorhersagen für neue Daten abzugeben, die von gleichen Populationen erzeugt werden oder um die zugrundeliegenden Prozesse zu beschreiben (Hastie u. a., 2001, vgl. S. 2). Dabei stützt sich die Definition solcher Prozesse auf Merkmale (Features) der Daten, die verschiedene Zustände repräsentieren. Von einem Datenzustand kann andererseits wieder auf die Zustände der Merkmale geschlossen werden, wenn unterschiedliche Daten einer Population analysiert werden.

Statistik und maschinelles Lernen mit Text

Die Verarbeitung von Text mit Werkzeugen des maschinellen Lernens funktioniert nur, wenn die Dokumente in einer Form repräsentiert sind, die für die Prozesse geeignete Eingabedaten darstellen. Es hat sich durchgesetzt, das Vokabular eines Korpus auf den Bereich der natürlichen Zahlen abzubilden, was eine einfachere Verarbeitung erlaubt.

$$f\colon W \to V, w \mapsto v, v \in \mathbb{N} \tag{2.1}$$

oder

$$v = f(w) \tag{2.2}$$

In dieser Abbildung ist W die Menge der in einer Textkollektion (Korpus) verwendeten Types, wobei jedem Type eine natürliche Zahl zugeordnet wird. Die Types sind die unterscheidbaren Objekte in einem Korpus, wie beispielsweise die verschiedenen Wortarten. Im Gegensatz dazu sind Token die Instanzen der Wortarten in den Dokumenten (Manning, 1999, vgl. S. 22). Die Repräsentation der Worte eines Textes als natürliche Zahlen hat den Vorteil, dass die Dokumente nun als

- Folge natürlicher Zahlen $v_i = \{v_1, \ldots, v_n\}$ oder als

- Vektor der Dimension $\mathbb{R}^{|\mathbb{V}|}$ (Vektorraum Modell), wobei die Elemente beliebige reelle Zahlen annehmen, wie zum Beispiel Wortanzahl im Dokument, Wortwahrscheinlichkeiten oder Wortgewichte,

dargestellt werden können. In dieser Form ist es möglich, Dokumente in mathematisch statistische Verfahren einzubetten. Die Wahl der Darstellung hängt davon ab, welche Verfahren verwendet werden. Typischerweise wird für Verfahren, die eine Abhängigkeit oder Sequenz von Wörtern einbeziehen, eine Folge von Wörtern v_i generiert, während für bestimmte Klassifikatoren, bei der nur das verwendete Vokabular beachtet wird, eine vektorielle Darstellung gewählt wird. Der Ansatz, dass die Dokumente ohne Beachtung der Reihenfolge darstellt werden, ist unter dem Begriff „Bag of Words“ bekannt (Manning, 1999, vgl. S. 237).

Überwachtes und unüberwachtes Lernen

Beim Einsatz der Verfahren aus dem Machine-Learning zeigt sich eine Verbindung zu den Methoden der Inhaltsanalyse. Es können deduktive und induktive Ansätze unterschieden werden, die allerdings als überwachte und nicht überwachte Verfahren (supervised, non-supervised) bekannt sind (Hastie u. a., 2001, vgl. S. 3). Jedes Objekt (z.B. ein Dokument) in einer Kollektion von Daten (z.B. ein Korpus) weist unterschiedliche Ausprägungen seiner Merkmale auf. Die Trainingsdaten oder -objekte $\mathbf{x} = (x_1, \ldots, x_p)$ und die Features der einzelnen Instanzen i der Trainingsdaten $x_i = (x_{i1}, \ldots, x_{ip})$, stellen für einen Algorithmus die Input Variable dar. Im Fall des überwachten Lernens enthalten die Trainingsdaten zusätzlich zu den Features eine Output Variable $\mathbf{y} = (y_1, \ldots, y_m)$, zum Beispiel eine Dokumentklasse oder eine Kategorie, und der Algorithmus kann angeleitet werden, die Ausprägungen der Features mit der Output Variable zu assoziieren. Den Fehler, den ein Algorithmus bei der Bestimmung der Output Variable unter Einbeziehung der Input Variablen macht, wird durch eine Verlustfunktion repräsentiert. In diesem Fall ist die Dichtefunktion $p(\mathbf{y}|\mathbf{x})$ von Interesse, deren Bestimmung notwendig ist, um optimale Modellparameter zu bestimmen. Beim unüberwachten Lernen existiert die Output Variable nicht und die Features werden genutzt, um die Daten und deren Varianzstruktur bzw. deren zugrunde liegende Dichtefunktion $p(\mathbf{x})$ zu schätzen. In anderen Fällen wird versucht, Regionen im Wertbereich von $\mathbf{x}$ zu finden, in denen die Varianz der Daten in diesem Bereich gering ist (Hastie u. a., 2001, vgl. S. 438). Dies ist beispielsweise beim Clustern der Fall. Im Vergleich mit der Inhaltsanalyse können die überwachten Verfahren als hypothesengeleitetes Vorgehen verstanden werden, da eine Trainingsmenge $\mathbf{x}$ manuell zusammengestellt wird. Diese Menge $\mathbf{x}$, mit je einem oder mehreren

Outputs $\mathbf{y}$, repräsentiert in diesem Fall eine Hypothese, was als deduktive Arbeitsweise angesehen wird. Im Gegensatz dazu schätzen die unüberwachten Verfahren die Verteilungs- und Dichtefunktionen $p(\mathbf{x})$ und stellen dadurch Strukturen in den Daten heraus. Im Fall von Textdaten, die durch das große Vokabular als Vektor mit sehr vielen Dimensionen dargestellt werden, wird die Struktur der Dichtefunktionen herangezogen, um Bereiche zu finden, welche die Daten vereinfacht darstellen (Hastie u. a., 2001, vgl. S. 438). Dies wird genutzt, um aus den Dichtefunktionen induktive Schlüsse zu ziehen und Hypothesen über die Daten zu generieren. Während durch die Verlustfunktion beim überwachten Lernen bestimmt werden kann, welche Fehler das System beim Schätzen des Outputs anhand der Daten des Prozesses macht, so ist dies beim unüberwachten Lernen nicht möglich. Die inferierten Strukturen können nicht ohne Weiteres beurteilt werden. Hier muss intuitiv und manuell kontrolliert werden, ob die bestimmten Dichten und Strukturen valide sind. Es können zwar quantitative Kennzahlen bestimmt werden, wie beispielsweise eine Varianz innerhalb verschiedener Bereiche der Dichtefunktion. Diese spiegeln aber nicht unbedingt die repräsentative oder interpretative Qualität des Ergebnisses wider.

Im Fall von Textsammlungen kann die Datenmenge als sogenannte Dokument-Term-Matrix beschrieben werden. Jedes Dokument ist als Vektor über das gesamte Vokabular ausgedrückt. In den entsprechenden Elementen des Vektors werden für alle Wörter in einem Dokument Häufigkeiten oder Gewichte eingetragen. Diese Matrix dient als numerische Repräsentation und Eingabe $\mathbf{x}$ der Textkollektion für statistische Verfahren.

Information Retrieval und explorative Suche

Der Zugriff auf große Textkollektionen wird durch Verfahren aus dem Machine-Learning unterstützt. Durch Verfahren des Information Retrieval (IR), die über Vektorraum-basiertes Klassifizieren, Clustering oder die Bestimmung latenter Variablen Zusammenhänge in Kollektionen definieren, werden reine schlüsselwortabhängige Suchen verbessert (Stock, 2007, vgl. S. 334 ff.). Eine Anfrage (Query) an das System setzt voraus, dass die Anfrageparameter in der Datenbank zu finden sind und Ergebnisse liefern (White u. Roth, 2009, vgl. S. 10). Um die Anfrageparameter zu definieren, muss die Kollektion oder das Korpus einigermaßen bekannt sein. Fehlen Kenntnisse über die Inhalte in der Datenbank, so kann es sein, dass nur die Objekte gefunden werden, deren Existenz schon bekannt ist. Um die Struktur und die Eigenschaften großer Datenmengen zu untersuchen, sodass ein tieferes Verständnis der Daten entwicklt werden kann, definierte Tuckey den Begriff „Exploratory Data Analysis“ (Tukey,

1977). Er betont darin, dass Methoden und Visualisierungen definiert werden müssen, die die Feature-Dichten, Objektgruppen und Hauptkomponenten der Daten visuell darstellen können. Mit diesen Methoden ist es möglich, Hypothesen und Annahmen über analysierte Daten zu generieren. Ferner können Gruppen und Abnormalitäten in den Daten durchsucht werden, um den Analyseprozess zu verbessern. Die Ansätze der explorativen Datenanalyse sind ursprünglich gedacht, um Daten zu verstehen und um Modelle für Daten zu verbessern. In Dokumentkollektionen müssen hingegen Arbeitsabläufe unterstützt werden, die dem Analysten helfen, naive Anfragen so umzugestalten, dass die Anfragen auf die Daten und die Datenmodelle passen. Die Aufgabe, eine Anfrage so zu erweitern oder zu verändern, dass ein Informationsbedarf an einer Textkollektion gedeckt werden kann, wird explorative Suche oder „Explorative Search“ genannt (Marchionini, 2006; White u. Roth, 2009). Das Ziel der explorativen Suche ist es, mit einer initialen Anfrage so unterstützt zu werden, dass die anfänglichen Annahmen anhand der Daten überarbeitet werden können und die Suche nach besseren Anfragen bzgl. eines Informationsinteresses unterstützt wird. Die explorative Suche in Dokumentkollektionen oder -korpora ist in zwei Phasen unterteilt. Die erste ist das explorative Browsing, bei dem die Nutzer eines Systems unterstützt werden, relevante Dokumente aufzufinden, die evtl. durch eine reine Schlüsselwortsuche nicht aufgefunden würden. Dabei können vorbereitete Klassifikationen, Metadaten oder Clusterprozesse helfen, deren Ergebnisse in der Datenbank vorliegen. Dieser Prozess kann die Nutzer soweit unterstützen, bis eine Dokumentmenge identifiziert ist oder eine Suchstrategie gefunden ist, die dem Informationsinteresse gerecht wird (White u. Roth, 2009, vgl. S. 20). Nach der Identifikation relevanter Objekte wird die Phase der fokussierten Suche definiert, sodass anhand der sicher gefundenen Dokumentmenge Informationen exploriert werden, welche die eigentliche Fragestellung beantworten können. In Abbildung 2.2 sind die Phasen dargestellt und es wird deutlich, dass die explorative Suche definiert worden ist, um Unschärfen in der Suchstrategie zu beseitigen und die Nutzer bei der Beseitigung der Unschärfe zu unterstützen. Systeme, die Konzepte der explorativen Suche nutzen, müssen Methoden verschiedener Disziplinen implementieren. Für die Untersuchung des Zusammenspiels aus Visualisierung, Interaktion, Datenanalyse, Datenmanagement und Statistik wurde ein interdisziplinärer Forschungsansatz, als Visual Analytics bezeichnet, entwickelt (Keim, 2010; Thomas u. Cook, 2005). Die Verfahren unterstützen die Nutzer bei Analysen, beim Schlussfolgern und beim Erkennen von Strukturen und Wissen in großen Datenmengen. Das Paradigma, was dem Visual Analytics Ansatz zugrunde liegt, besteht aus den Bestandteilen Daten, Visualisierung, Datenmodelle und Wissen. Insbesondere die

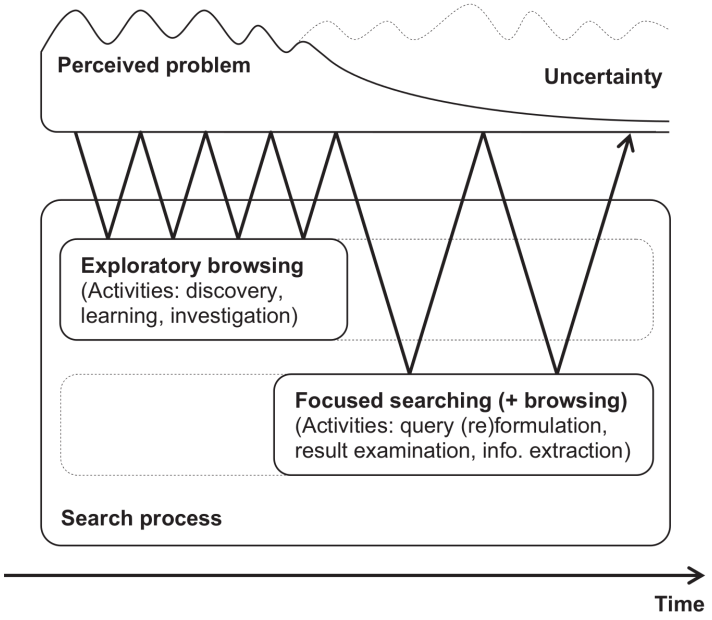

Abbildung 2.2: Ablauf einer explorativen Suche. Entnommen aus: White u. Roth (2009, vgl. S. 19).

Komponenten Visualisierung und Datenmodelle sind miteinander verzahnt, sodass durch die Datenmodelle bessere Visualisierungen der Daten generiert werden können. Andererseits soll es durch die visuelle Interaktion möglich sein, Modelle zu revidieren, um den Analyseprozess weiter zu verbessern oder zu fokussieren. [9] Im Gegensatz zu einem reinen Information Retrieval Ansatz erweitern die Methoden der explorativen Suche, und im erweiterten Sinn der Visual Analytics, den Analyseprozess um visuelle und datentechnische Ansätze. Innerhalb unbekannter Datenmengen ist es möglich, Strukturen zu identifizieren und die Modelle, Hypothesen und Anfragen zu optimieren. In reinen Retrieval Systemen können Hypothesen anhand von Anfragen geprüft werden. Sind allerdings keine Daten mit den operationalisierten Anfrageparametern auffindbar, fehlen Verfahren, um die Hypothesen umzugestalten oder bei der Anpassung und Suche helfen. Im Sinne der Inhaltsanalyse bilden diese unterschiedlichen Ansätze deduktive und induktive Arbeitsweisen, wie in Abschnitt 2.1.2 besprochen, ab. Ist bereits eine Hypothese so formuliert, dass sie direkt an einem Textdatenbestand getestet werden kann, so reichen herkömmliche Retrieval- oder Klassifikationsmodelle für die Arbeit aus. Muss der Datensatz oder die Dokumentmenge erst analysiert und operationalisierbare Hypothesen definiert werden, so empfiehlt sich bei der Inhalts-

[9] In der Literatur ist dieser Begriff als „Active-Learning“ bekannt (Settles, 2012).

analyse in großen Dokumentmengen ein Ansatz über die explorative Suche für die Definition der Hypothesen.

2.3 Zusammenfassung

Die linguistischen Theorien, um Themen in Texten zu erfassen, basieren alle auf der strukturellen, semantischen Analyse der analysierten Dokumente. Die Theorien unterscheiden grob zwischen bereits bekannten und neu eingeführten Konzepten oder Termen auf einer Satz- oder Dokumentebene. Die bekannten Elemente und Konzepte innerhalb einer Textstruktur, wie der Themenkern, Makropropositionen, ein Faktor bei der Fokustheorie, das Thema in der Thema-Rhema Gliederung oder die Erfassung von Nominalgruppen definieren den hauptsächlichen Zusammenhang eines Textes oder einer Dokumentsammlung. Die in der Linguistik entwickelten Strukturen für die formale Beschreibung eines Themas definieren und formalisieren im weitesten Sinn den Begriff Sinnkomplex, wie ihn Luhmann (1979, vgl. S. 13) beschreibt. Luhmann führt in seiner Definition aus, dass diese Sinnkomplexe entwicklungsfähig und unbestimmt sind. Die von den linguistischen Betrachtungen besprochene Entfaltung eines Themas deckt sich mit dieser Sichtweise. Auf der Grundlage dieser Erkenntnis kann festgehalten werden, dass eine automatische Themenextraktion die bekannten und die entfaltenden Objekte in den Texten so strukturieren muss, dass der Themengehalt einer Textkollektion erkennbar repräsentiert wird.[10] Es soll nicht auf die Trennung der Sprachintention ankommen, sondern um die Beurteilung eines gemeinsamen Sinns der Dokumente. Die verschiedenen linguistischen Theorien weisen die Gemeinsamkeit auf, dass die thematischen Strukturen durch mehr oder weniger strukturierte Wortgruppen repräsentiert sind. Die Thema-Rhema Gliederung und die Ableitung von Makropropositionen formalisieren die Darstellung der Strukturen am stärksten. Die stark strukturierten Themendarstellungen sind nach näherer Betrachtung aber subjektiver und komplizierter zu erstellen. Die Darstellung von Wort- und Nominalgruppen oder Faktoren lässt sich wesentlich einfacher nachvollziehen und abstrahieren und reduzieren die Informationen in Texten. Thematische Übereinstimmungen in den Texten können durch die Festlegung abstrahierter Wortgruppen dargestellt werden. Je mehr allerdings von der thematischen Struktur genommen wird, desto weniger kann die Entfaltung des Themas verstanden werden. Die im Folgenden untersuch-

[10] Auch die Terminologie der dynamischen und weniger dynamischen Textbestandteile beschreibt dieses Konzept.

ten Methoden müssen danach beurteilt werden, wie analysierte Themen abstrahiert werden und wie das Thema als Resultat einer Analyse dargestellt werden kann.

Die Idee, Themen über die distributionellen Eigenschaften der Texte zu erfassen, legt nahe, dass thematische Strukturen in den Texten manifest sein müssen. Demnach obliegt es der Analysemethode, die Strukturen für die Themenabstraktion zu erfassen. Dies führt zu einer Diskussion, ob die Strukturen deduktiv-theoretisch vorgegeben werden sollen, um deren Existenz in den Texten zu prüfen, oder ob die Strukturen aus den Texten extrahiert werden, um die Themenextraktion induktiv zu gestalten. Bei großen Textkollektionen sind die Ausprägungen der darin enthaltenen Themen nicht alle bekannt oder es ist sehr schwer dies zu erfassen. Es können natürlich Theorien vorhanden sein, die eine enge theoretische Definition eines Themas erfordern. Daraus resultiert eine fokussierte Suche nach konkreten Textinstanzen. Die strukturelle Gliederung des Korpus in eine Themenaufteilung ist in diesem Fall nicht nötig. Durch die enge Definition wird der Suchraum eingegrenzt und die Analyse findet anhand der Dokumente statt, die exakt zur Themendefinition passen. Deshalb werden zur Themenidentifikation im Folgenden unüberwachte automatische Methoden untersucht, da die automatische thematische Untersuchung großer Textmengen ohne Vorwissen auskommen muss und die Suche nach vorgegebenen Themen mit überwachten Methoden bereits untersucht wurde (Scharkow, 2012). Diese Vorbedingung soll deshalb gelten, da die Ausprägungen der Themen in großen Textmengen nie vollständig im Vorhinein festgelegt werden können. Weiterhin sind unüberwachte Verfahren insbesondere wichtig, um in der Vorphase einer deduktiven Untersuchung, die Kategorien zu entwickeln und deren Bestand in den Quellen zu überprüfen. Die Analyse von wachsenden Textkollektionen erfordert im besonderen Maße den Einsatz unüberwachter Methoden, da sich die Themenstruktur jederzeit ändern kann und nicht vorhersehbar manifestiert ist. Mit dem Einsatz unüberwachter Methoden, die mit Hilfe der distributionellen Semantik einer Textkollektion Themen auffinden, werden im Folgenden die Ziele

- Aggregation von Dokumenten, die eine thematische Struktur zu einer bestimmten Abstraktion teilen,
- Aufbereitung der Strukturinformationen und der Wortverwendungszusammenhänge der Themen in einer Textkollektion, sodass eine semantische Beurteilung und Interpretation gemäß linguistischer Themendarstellungen erfolgen kann,
- Auffinden und Bearbeitung mehrerer Themen,
- Erschließung noch nicht offensichtlicher Themen aus unbekannten oder nicht erschlossen Quellen und

- Anwendung der Methode der explorativen Suche zur Identifikation von Themen, die anhand einer Hypothese oder Theorie generiert werden, verfolgt.

Anhand der Methoden der automatischen Sprachverarbeitung und den theoretischen Konzepten zur Bestimmung von Themen und Themenzyklen soll eine Systematik entwickelt werden, um innerhalb von Themen

- Personen,
- Orte,
- Ereignisse,
- Wortverwendungen und
- Themenzyklen

als Bestandteil eines Themas zu ermitteln und zu aggregieren. Zu diesem Zweck sollen die Verfahren der automatischen Themenanalyse mit anderen Verfahren, wie der Zeitreihenanalyse, der Kookkurrenzanalyse und der automatischen Eigennamenserkennung kombiniert werden. Durch die angesprochenen ökonomischen Probleme der Inhaltsanalysen und der aufwändigen Kategorisierung, wird das Potential dieser integrieren Anwendung verschiedener Verfahren des Text-Mining für die unüberwachte Codierung von Themen in großen Textkollektion untersucht. Es muss ein methodisches und technisches Vorgehen erarbeitet werden, um automatisierte Themenanalysen in Textkollektionen prospektiv, retrospektiv, synchron und diachron durchzuführen. Basierend auf der Untersuchung geeigneter Verfahren und einer praktischen Erprobung kann eine Einschätzung darüber abgegeben werden, wie sich die verschiedenen Untersuchungsmethoden mit automatischen Verfahren und einer geeigneten Datenhaltung abbilden lassen.

Die Themenstrukturerkennung kann für die Selektionen und Filterungen in Textsammlungen genutzt werden, sodass Dokumentmengen anhand der Themenstruktur eingeschränkt werden können. Ein weiterer Aspekt der folgenden Ausführungen soll deshalb die Darstellung der analysierten Themenstrukturen sein. Die Ergebnisse automatisierter Methoden müssen verständlich und interpretierbar dargestellt werden können. Die Ergebnisse der Verfahren müssen anhand inhaltsanalytischer Anforderungen dargestellt werden, um geeignete Methoden zur explorativen Arbeit anzubieten. Diese Möglichkeiten werden in Kapitel 4 genauer untersucht.

2.4 Konkretisierung der Forschungsfragen

Die distributionelle Semantik ist sehr hilfreich bei der Beurteilung von Spracheigenschaften. Häufig verwendete Worte können identifiziert werden, um so die Strukturen und Semantik einer Sprache verstehen. Über die Verteilungen der Wörter innerhalb der Dokumente kann beobachtet werden, welches Vokabular zwischen einzelnen Dokumenten geteilt wird und welche Wortstrukturen individuell in den Dokumenten verwendet werden. So muss es nach den linguistischen Themendefinitionen Bereiche der Wortverteilung $p(\mathbf{x})$ geben, die sich in den Dokumenten zu einem Thema überdecken, wenn das Thema bekannte Konzepte darstellt über die etwas neues gesagt wird. Wenn sich mehrere Dokumente, wie bei Nachrichtentexten, mit einem Thema beschäftigen, so müssen diese Dokumente einen Überdeckungsgrad der Wortverteilungen besitzen. Dieser kann durch automatische Methoden genutzt werden, um eine thematische Verwandtschaft der Dokumente zu bestimmen. Im folgenden Kapitel wird deshalb eine Auswahl an Methoden vorgeschlagen, die die distributionelle Semantik der Texte nutzen, um die Themenstrukturen zu bestimmen. Die vorgeschlagenen Methoden werden dahingehend untersucht, ob die Darstellungsmächtigkeit, Struktur und semantische Verständlichkeit der linguistischen Thementheorien abgebildet ist. Für die Untersuchung und Beurteilung der vorgeschlagenen Verfahren werden deshalb folgende Forschungsfragen formuliert, die unter Berücksichtigung der theoretischen Themenvorstellungen und den Anforderungen der inhaltsanalytischen Themenanalysen erstellt werden. Im Wesentlichen reflektieren die Fragen auch die Anforderungen, die an eine automatische Themenanalyse gestellt werden, wenn die Methoden zu kommunikationswissenschaftlichen und linguistischen Grundlagen und Methoden anschlussfähig sein sollen.

- **F1: Lassen sich die Sichtweisen und Methoden der Inhaltsanalyse und speziell der Themenanalyse in der Anwendung computergestützter Verfahren wiederfinden? Wie weit lassen sich Verfahren, die den Ansprüchen der Themenanalyse genügen, automatisieren?**
  Die Beantwortung dieser Frage soll vor allem klären, wie anschlussfähig die Methoden der automatischen Sprachverarbeitung und des Text-Mining an die Inhaltsanalyse sind und welchen Grad die Automatisierbarkeit erreichen kann. Es soll ein methodisches Vorgehen erarbeitet werden, welches nötige Vorbereitungen, Daten und sequenzielle Analyseschritte dokumentiert.

- **F2: Haben geeignete Verfahren für die automatische Themenanalyse einen quantitativen oder qualitativen Charakter?**

Diese Frage zielt insbesondere auf das Potential der automatischen Verfahren ab, Themenstrukturen neben der reinen Deskription auch interpretierbar zu machen. Dafür kann es unter Anderem nötig sein, verschiedene Skalenniveaus aus den Analysen ableiten zu können. Die Ableitung nominaler Skalen in Form von benannten Kategorien wäre ein Beispiel für die qualitative Sichtweise auf Dokumentmengen und Themen. Beurteilungen der Berichterstattungsmengen zu Themen über die Rationalskala wären dagegen ein Beispiel für quantitative Sichtweisen.

- **F3: Haben geeignete Verfahren für die automatische Themenanalyse einen deduktiven oder induktiven Charakter?**
  Die Idee, diese Frage zu stellen, liegt darin begründet, dass die Verfahren dahingehend beurteilt werden müssen, ob sie zur reinen Exploration dienen oder für das Testen von Theorien geeignet sind.

- **F4: Wie valide und reliabel sind automatische Verfahren in der Themenanalyse einsetzbar?**
  Die automatischen Verfahren müssen im Hinblick auf ein linguistisches Themenverständnis geprüft werden. Nur wenn die Verfahren Ergebnisse liefern, die Interpretationen und Darstellungen im Sinne dieser Themenkonzepte liefern, können sie sinnvoll interpretiert werden und sind valide Messverfahren. Zusätzlich ist es wichtig, dass die Verfahren stabil arbeiten und die Ergebnisse nachvollziehbar und reproduzierbar sind.

- **F5: Wie lassen sich die Verfahren nutzen, um die analysierten Themen in Dokumentmengen und Zeitreihen zu überführen, sodass Querschnittanalysen, die Analyse von Nachrichtenfaktoren und -werten und die Bestimmung von Themenzyklen möglich werden?**
  Die Ergebnisse der Verfahren müssen nicht nur valide, reliabel, deskriptiv oder qualitativ sein. Die Weiterverwertung der Analyseergebnisse und deren Einbettung in kommunikationswissenschaftliche Betrachtungen erlaubt vielfältige Schlüsse auf Sender und Empfänger von Nachrichten oder auf die inhaltlichen Filterfunktionen der Übertragungswege. Hierfür müssen die reinen Themen oder deren kategorialen Beschreibungen weiterverarbeitet werden, um eine geeignete Operationalisierung der Theorien und Betrachtungsweisen zu schaffen. In Abschnitt 2.2 werden unterschiedliche Quellformate beschrieben. Eine Untersuchung muss zeigen, ob die Auswertungen der Themen für verschiedene Quellen

vergleichbar sind und so Schlüsse und Zusammenhänge zwischen verschiedenen Quellen gefunden werden können.

- **F6: Welche Aspekte der Datenhaltung und Datenverarbeitung müssen bei der Anwendung computergestützter Methoden beachtet werden?**

  Wie weiter oben erläutert wird, können Analysen retrospektiv, prospektiv, synchron oder diachron durchgeführt werden. Diese Unterscheidungen haben für die Datenhaltung und Datenverarbeitung direkte Konsequenzen. So müssen Datenbanken oder Datensammlungen zeitliche Aspekte repräsentieren oder offen für neue Daten sein. Die Verfahren für die Themenanalyse müssen für unterschiedliche Untersuchungsstrategien anwendbar sein. Im Zuge der folgenden Untersuchungen sollen die Konsequenzen für die Datenhaltung und -verarbeitung deutlicher herausgearbeitet und dokumentiert werden.

  Diese Frage untersucht die Hypothese, dass in automatischen Verfahren, welche die distributionelle Semantik von Texten erfassen, die dokumentübergreifenden Themenstrukturen in mehreren Dokumenten in einen Zusammenhang gebracht werden können. Nur so ist es möglich, Dokumente einem Thema zuzuordnen und zu quantifizieren. Mit Hilfe der Zeitstempel sind Querschnittanalysen möglich, die ein Thema bezüglich der öffentlichen Aufmerksamkeit valide abbilden. Thematisch zusammenhängende Dokumente müssen nach Informationen durchsucht werden können, um beispielsweise Akteursstrukturen zu erfassen. In diesem Sinne soll die Antwort auf die Frage aus einer Untersuchung der Verfahren abgeleitet werden. Diese soll insbesondere zeigen, wie die Eignung der Verfahren für die Erstellung weiterführender Analysen der Themen, wie die Nachrichtenfaktoren oder Themenzyklen, zu beurteilen ist.

Die komplexeren Thementheorien, wie die Rhema-Thema Gliederung oder die Verwendung von Makropropositionen, erfordern eine starke subjektive und individuelle Leistung der Analysten und sind schwer intersubjektiv nachvollziehbar. Somit wird die Anwendung automatisierter Verfahren im Rahmen einiger Theorien erfolgreicher und in anderen weniger erfolgreich sein. Die Untersuchung der Fragen F1 und F4 soll die Anschlussfähigkeit der automatischen Verfahren an die Methode der Inhaltsanalyse und die linguistischen Betrachtungsweisen von Themen und deren praktische Eignung überprüfen. Zusätzlich soll untersucht werden, wie die Verfahren Arbeitsabläufe vereinfachen und automatisieren, sodass große Dokumentmengen erschlossen werden können. Wenn es möglich ist, aus Themenabstraktionen ein tieferes Verständnis des

Themas herzuleiten, stellt die automatische Erfassung von Textthemen auch einen Beitrag zur Erhöhung der Reliabilität bei der Themenerfassung für große Dokumentsammlungen dar. In F5 soll dagegen geklärt werden, wie die Ergebnisse und Daten einer automatischen Themenanalyse weiterverwendet werden können, um weitere Informationen und Eigenschaften der Themen nachträglich analysierbar zu machen. Um eine Aussage zu den Forschungsfragen treffen zu können, werden geeignete Verfahren in Kapitel 3 vorgestellt und ihre Eigenschaften, Ergebnisse und mögliche Auswertungen werden untersucht und dargestellt. In Kapitel 3 schließt sich eine Anwendung der Verfahren mit implementierten Software-Werkzeuge an, um mit einer pragmatischen Perspektive Antworten auf die Forschungsfragen zu geben.

Kapitel 3

Algorithmen und Methoden für die automatische Themenanalyse

Die im letzten Kapitel besprochenen Grundlagen dienen zur Systematisierung und Einordnung der Anforderungen an eine Themenidentifikation innerhalb von Textkollektionen. Es stellt sich heraus, dass für die Themenbeobachtung und Identifikation vor allem unüberwachte Verfahren interessant sind, da die Bildung von Hypothesen und Trainingsdaten für große Textmengen auf unvollständigem Wissen basiert. In den nun folgenden Abschnitten werden Verfahren vorgestellt und bewertet, die den Anforderungen, die in Kapitel 2 definiert werden, genügen. Die Verfahren werden an dieser Stelle in ihren Eigenschaften erklärt und kritisiert. Dadurch wird der Umgang mit den Verfahren und deren Nutzen hinreichend für den Anwendungsfall untersucht. Anhand der Ergebnisse wird gezeigt, welche Möglichkeiten die Verfahren aufzeigen, um Textthemen zu erfassen und zu verstehen und wie Anteile der Themen, auch diachron, an einem Korpus gemessen werden können. Anhand der Themen wird bestimmt, wie sich die Wortverwendung eines Themas, hinsichtlich theoretischer Vorgaben der Linguistik und der Inhaltsanalyse, anhand distributioneller Semantik erklären und zusammenfassen lässt, um eine valide Interpretation der Bedeutungen gewährleisten zu können.

Im Hinblick auf den dargestellten Zusammenhang zwischen einzelnen Ereignissen, Themen, Stories oder der Wortverwendung, wird untersucht, wie unterschiedliche Verfahren in diese Sichtweisen einzuordnen sind und wie unterschiedliche Abstraktionsgrade thematischer Betrachtungen analysiert werden können. Unter einem Abstraktionsgrad wird hier der Umfang der Informationsreduktion verstanden, die vorgenommen werden muss, um die Inhalte eines Themas zusammenzufassen. Werden beispielsweise Einzelereignisse isoliert dargestellt, so sind dort detaillierte Einzelinfor-

mationen enthalten. Wird ein Ereignis in einen umfassenderen Bericht integriert, so müssen mehr Informationen ausgeblendet werden, um eine Zusammenfassung aller Inhalte zu liefern. Soll ein Zusammenhang in eine Rubrik eingeordnet werden, treten Ereignisse ganz in den Hintergrund und die thematische Darstellung reduziert sich auf Eigenschaften der Rubrik. Da diese Unterschiede der Abstraktionsgrade in thematischen Untersuchungen beachtet werden müssen, wird gezeigt, wie Zusammenfassungen auf unterschiedlichen Abstraktionsgraden möglich sind. Wie im Abschnitt 2.2.2 besprochen, konzentrieren sich die Verfahren auf unüberwachte Verfahren des maschinellen Lernens, da aufwändige Definitionen von Regeln und Trainingsmengen nicht praktikabel für Daten sind, deren Inhalte nicht ausreichend bekannt sind. Die Ausprägung bestimmter Kategorien kann nicht vollständig und gemäß der vorliegenden Daten in großen Korpora bestimmt werden. Eine weitere Abgrenzung ist nötig. Zu den etablierten clusternden und unüberwachten Verfahren gehören das k-Means Verfahren und agglomerative hierarchische Clusterverfahren(Heyer, 2006, vgl. S. 198 ff.). Dimensionsreduzierende Verfahren wie „Latent Semantic Indexing“ (LSI bzw. LSA) oder die „Principal Component Analysis“ (PCA) eistieren, die ähnliche Verteilungsstrukturen in Dokumenten abbilden(Hofmann, 1999). Eine Untersuchung und Evaluierung dieser Verfahren wird in diesem Kapitel nicht durchgeführt, da die hier vorgestellten Verfahren optimierte und verbesserte Varianten bzw. Neuentwicklungen dieser Verfahren darstellen. Die hierarchischen agglomerativen Clusterverfahren sind nicht geeignet, da die Dokumentmenge nachträglich nicht verändert werden kann oder verschiedene Stufen in der Clusterhierarchie zwischen unterschiedlichen Dokumentmengen nicht vergleichbar sind. Während der Diskussion der vorgestellten Verfahren werden die Bezüge zu den Standards deutlicher gemacht.

3.1 Topic Detection and Tracking

Unter dem Namen „Topic Detection and Tracking“ (TDT) ist ein Forschungsprojekt für die ereignisorientierte Organisation von Nachrichten (Allan, 2002, vgl. S. 1 ff.) zusammengefasst. Auch audiovisuelle Inhalte und deren Organisation und Verarbeitung zählen dazu. Die Aufgaben, welche innerhalb der Forschungslinie bearbeitet werden, gliedern sich in

- Story Segmentation, die Zerlegung eines Nachrichtenstroms[1] oder einer Textsammlung in diskrete Einzelnachrichten,

[1] Unter einem Nachrichtenstrom wird eine sequenzielle Abfolge von Nachrichten verstanden.

- First Story Detection, die Detektion neuer unbekannter Nachrichten,
- Cluster Detection, die Gruppierung aller Nachrichten anhand der Inhalte,
- Tracking, Auffinden ähnlicher Nachrichten anhand von Beispielen in einem Nachrichtenstrom und
- Story Link Detection, der Untersuchung von Methoden, die feststellen, ob zwei Nachrichten ähnliche Themen oder Ereignisse besprechen.

Jede der Aufgaben bildet im TDT-Zusammenhang eine eigene Forschungsrichtung mit eigenen Evaluierungsrichtlinien. Für jeden Teilbereich existieren dadurch mittlerweile Ansätze, die jeweils für eine der oben genannten Aufgaben entworfen und evaluiert wurden. So beschäftigt sich der Tracking-Ansatz mit Lösungen, die anhand einer Textkollektion und Beispieldaten für existierende Ereignisse in der Lage sind, den vorgegebenen Ereignissen Dokumente zuzuordnen. In diesem Fall wird ein Thema als Sammlung von Einzeldokumenten oder „Stories“ bezeichnet, die sich auf ein Thema beziehen (Allan, 2002, vgl. S. 25). Im Gegensatz dazu steht bei der Link Detection das Ziel im Vordergrund, Stories zu vergleichen und zu entscheiden, ob diese Stories zu einem Thema zugehörig, also verbunden, sind. Bei der Detection kommt es darauf an, dass aus einer Textkollektion Themen ohne ein Vorwissen erkannt werden. Das Ziel dieser Einzeluntersuchungen ist, dass in der Kombination aller Ergebnisse ein Gesamtsystem zur Themenerkennung und -strukturierung entwickelt werden kann. In aller Kürze sollen an dieser Stelle nur einige daraus entstandene Ansätze dokumentiert sein. Für die Aufgaben Detection und Tracking wurden beispielsweise probabilistische Ansätze, Sprachmodellierung, Clustering oder Verfahrenskombinationen erprobt (Allan, 2002). Die Ergebnisse wurden in einem Abschlussbericht und einem Buch zusammengefasst (Allan u. a., 1998; Allan, 2002). In einem weiteren Aufsatz gehen die Autoren allerdings noch einmal auf die praktische Umsetzung und Praxisprobleme ein, die bei großen und realen Datenmengen abseits der Testdaten auftreten (Allan u. a., 2005). Die Autoren schildern darin ein robustes Gesamtsystem unter Gesichtspunkten der Anwendung und der Daten.

Sollen die Konzepte in ein Gesamtsystem integriert werden, zeigt sich, dass sich die einzelnen Disziplinen gegenseitig bedingen. Das Tracking von gleichen Nachrichten wäre ohne den Einsatz einer zuverlässigen Story Link Detection Strategie nicht möglich. Dies gilt insbesondere für die Bildung von Nachrichtenclustern. Für die angestrebte Messung und Zusammenfassung von Themenstrukturen werden, für diese Arbeit, nicht alle Teile aus diesem Forschungsbereich nützlich sein. Die Story Segmentierung soll an dieser Stelle nicht weiter ausgeführt werden, da in den untersuchten

Textkollektion davon ausgegangen wird, dass bereits segmentierte Dokumente mit jeweils homogenem Inhalt vorliegen. Wichtig ist es für die hier angestrebte Untersuchung, dass einzelne Texte unterschieden werden können und über Cluster- oder Tracking-Mechanismen gruppiert werden, sodass eine Grundstruktur unterschiedlicher Themen bzw. „Stories“ dargestellt und mit quantitativen Maßen aggregiert werden kann.

Innerhalb des Forschungsbereichs TDT wird eine Trennung unterschiedlicher Granularitäten bzw. Abstraktionen von Nachrichteninhalten festgelegt. Es werden grundsätzlich drei Arten zu messender oder beschreibender Zusammenhänge genannt, auf die sich die Arbeiten beziehen.

- Ereignisse (Events) sind „something that happens at some sprecific time and place [...]“ (Allan, 2002, vgl. S. 19).[2]
- Stories (Artikel) werden nach einer Definition des Linguistic Data Consortium (LDC) als „a topically cohesive segment of news that includes two or more declarative independent clauses about a single event“ eingeführt (Allan, 2002, vgl. S. 18).[3]
- Themen (Topics) werden als „a seminal event or activity, along with all directly related events and activities“ (Allan, 2002, vgl. S. 19) definiert.[4]

Anhand der Einordnung ist zu erkennen, dass die Verfahren stark auf die Eigenschaften von Ereignissen zugeschnitten sind und unterschiedliche Methoden vorhanden sein müssen, um auf den drei Ebenen zu aggregieren oder zu gruppieren. Diese ereignisorientierte Sichtweise scheint erst einmal nicht zu einer linguistischen Sichtweise zu passen, deren Augenmerk nach (Lötscher, 1987) auf mangelhaften Objekten im Text liegt. Bei genauerer Betrachtung sind aber die Ereignisse genau die mangelhaften Objekte, auf die im Text Bezug genommen wird und die genauer erklärt werden. Die aus den Verfahren ablesbaren Themen, stellen immer Ereignisse und Stories als Mangelobjekte in den Vordergrund. Damit sind sie für die Themenanalyse ereignisorientierter Textquellen geeignet. Vor allem wenn die Quelle viele Texte zu einzelnen Ereignissen enthält. Dies trifft vor allem für Nachrichtenkorpora zu.

[2] Etwas, dass zu einem bestimmten Zeitpunkt und Ort geschieht oder geschehen ist.

[3] Ein thematisch zusammenhängendes Segment von Nachrichten, die zwei oder mehr beschreibende und unabhängige Sätze über ein einzelnes Ereignis enthalten.

[4] Ein grundlegendes Ereignis oder eine Aktivität mit allen zugehörigen Ereignissen und Aktivitäten.

3.1.1 Clustermethode

Das in Allan u. a. (2005) und Stock (2007, vgl. S. 425 ff.) empfohlene System deckt die Anforderungen für die Bereiche Tracking, Detection, First Story und Link Detection ab. Der Ansatz liefert unter realen Bedingungen die beste Leistung bezüglich Genauigkeit und Rechenaufwand. Er besteht aus einer dem Relevance Feedback beziehungsweise dem Roccio-Algorithmus (Manning u. a., 2008, vgl. S. 178 f.) verwandten Methode. Jede zu analysierende Story bzw. jedes Dokument wird in einen Termvektor, wie im Abschnitt 2.2.2 beschrieben, überführt. Die Gewichtung der einzelnen Terme in einem Dokument erfolgt nach dem Schema

$$w_{t,s} = \frac{tf_{t,s} \cdot \log((0.5 + N)/df_t)}{\log(1.0 + N)}. \tag{3.1}$$

Dabei ist $tf_{t,s}$ die Häufigkeit eines Terms für ein Dokument bzw. Story s und df_t die Dokumentfrequenz eines Terms. Im Wesentlichen entspricht dieses Vorgehen einer Termgewichtung mit dem TF/IDF Maß (Manning, 1999, vgl. S. 543). Für jedes Dokument werden nun alle Termgewichte absteigend sortiert und eine Unterauswahl über die Top 1000 Termgewichte eines Dokuments durchgeführt.[5] Auch bei langen Dokumenten ist der Inhalt damit kompakt und spezifisch repräsentiert. Mit einem Termvektor wird die Ähnlichkeit zu allen anderen Dokumenten, die dem System schon bekannt sind, mit dem Kosinusmaß berechnet.

$$sim(A, B) = \frac{\sum_t w_{t,A} \cdot w_{t,B}}{(\sum_t w_{t,A}^2 \sum_t w_{t,B}^2)^{0.5}} \tag{3.2}$$

Hier spiegelt w die Termgewichte eines Dokuments A oder B wieder. Für die Ähnlichkeit wird ein Schwellwert von 0.21 angegeben (Allan u. a., 2005). Erreicht kein Vergleich dieses Mindestmaß, so kann das Dokument keinem bekannten Dokument zugeordnet werden und es wird eine neue leere Dokumentmenge mit diesem Dokument angelegt. Wird der Schwellwert überschritten, so wird das Dokument einer Dokumentgruppe (Cluster) zugeordnet, welche das ähnlichste Dokument enthält.[6] Ist C also eine Dokumentgruppe oder ein Cluster, so wird dem Dokument A eine Cluster nach folgender Regel zugewiesen:

$$C(A) = C(\text{maxsim}B). \tag{3.3}$$

[5] Da in Nachrichtenkorpora Stories in der Regel einem Artikel entsprechen, wird die Bezeichnung Dokument verwendet.

[6] Dies entspricht einem Single-Link Clustering.

Kumaran u. Allan (2005) stellen fest, dass der reine Vergleich der Termvektoren nicht ausreicht. Durch Personen, Orte oder Organisationen, die Named Entities, die sich innerhalb verschiedener Stories oder in Ereignissen gleichen können, werden durch das Verfahren Ähnlichkeiten unterstellt, die nicht da sind. Deshalb trennen sie den Abstand der Dokumente in zwei Vergleiche auf. Die Named Entities werden als eigener Termvektor repräsentiert und die restlichen Terme in einem eigenen Vektor isoliert. Nur wenn beide Vergleiche einen Schwellwert überschreiten werden sie bekannten Clustern zugeordnet.

3.1.2 Anwendung

Um die Dokumente einer Dokumentkollektion im Sinne der Termvektor-Clusterung nach Formel 3.1 zu verarbeiten, müssen die Dokumente in Termvektoren überführt werden. Das Verfahren ist ein Bag-of-words Ansatz, sodass die Dokumente als Vektor über das Vokabular dargestellt werden müssen, wie es auf Seite 45 gezeigt wird. Die Dokumente müssen sequenziell, nach Zeitstempel, sortiert und in mehreren Stapeln (Batches) übergeben werden, sodass die Dokumente den Clustern zugeordnet werden können. Um aus einer Textkollektion, die statisch vorliegt und nicht einen kontinuierlichen Textstrom darstellt, Themencluster im diesem Sinne erzeugen zu können, muss jedes Dokument in sequenzieller Reihenfolge ausgelesen werden. Die Batches umfassen Dokumente, die zusammen in einer Sekunde, Stunde oder an einem Tag datiert werden. Im Detail vollzieht sich der Ablauf nach Prozedur 1, die im Anhang verzeichnet ist.

Auf diese Art und Weise können sowohl statische, diachrone Textkollektionen, als auch zeitlich, sequenziell erweiterbare Textkollektionen mit diesem Algorithmus verarbeitet werden. Auffällig ist jedoch, dass die Anzahl der Vergleiche linear mit den Dokumenten, die bereits verglichen wurden, wächst. Im Fall einer sehr großen Dokumentmenge wird der Aufwand ein Dokument einem Cluster zuzuordnen immer höher, da das letzte Dokument theoretisch mit allen Dokumenten verglichen werden muss. Aus diesem Grund ist es sinnvoll, die Menge der bereits verarbeiteten Dokumente stabil und klein zu halten. Der Algorithmus ist so zu erweitern, dass die Menge der bereits gesehenen Dokumente auf ein retrospektives Fenster begrenzt wird. Im Fall von Nachrichtentexten kann die Quelle beispielsweise in tageweisen Stapeln verarbeitet werden. Die Menge p kann auf die letzten n Tage, Monate oder Jahre beschränkt werden. Je nachdem, welche Quellen vorliegen, können die Horizonte so gewählt werden, dass die Berechnung effizient verläuft und Verbindungen ähnlicher thematischer Bezüge dennoch hergestellt werden können. Ausgehend von der Annahme, dass die

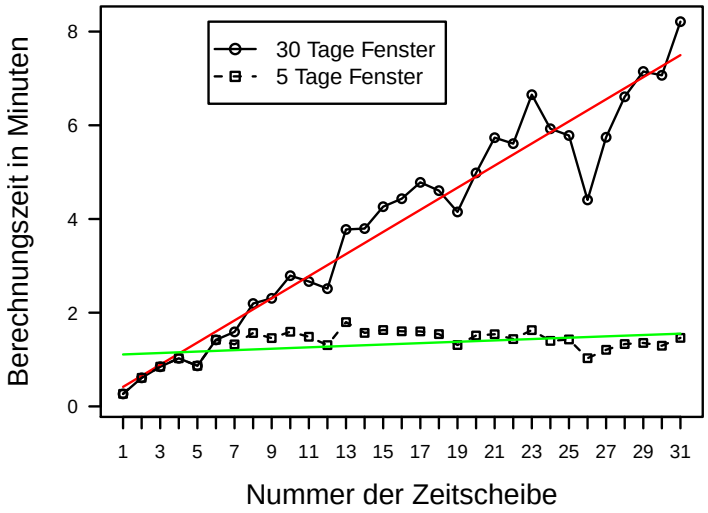

Abbildung 3.1: Laufzeitverhalten zur Berechnung der Batches für unterschiedliche retrospektive Horizonte über Tageszeitscheiben.

Bezüge, Ereignisse oder Themen innerhalb diachroner Textquellen über die Zeit an Bedeutung verlieren oder ersetzt werden (Allan u. a., 1998), kann diese Vereinfachung getroffen werden, da mit größerem Zeitabstand die Wahrscheinlichkeit sinkt, eine Verbindung zu aktuelleren Dokumenten zu finden. Durch den hohen Ereignis- und Story-Bezug dieses Verfahrens, ist eine Ähnlichkeit von Dokumenten innerhalb einer geringen Zeitspanne wahrscheinlicher. In Abbildung 3.1 ist ein Experiment dokumentiert, welches die Laufzeit des Verfahrens aufzeichnet. Es ist deutlich zu erkennen, dass eine Begrenzung des retrospektiven Horizonts für den Dokumentvergleich, die Laufzeit erheblich verringert. Vor allem bleibt die Berechnungszeit über die Zeitscheiben hinweg stabil, da die Menge der Vergleichsdokumente für die Zeitscheiben stabil gehalten wird.[7] Nur über eine solche Strategie oder aber über geeignete beiläufige Verfahren ist die Verarbeitung in großen Textmengen effizient.

Das Ergebnis dieses Verfahrens ist eine Menge von Clustern C, denen jeweils eine Menge von Dokumenten, D_C, zugeordnet ist. Innerhalb der Cluster ist jedes Dokument als Termvektor repräsentiert. Dieser Vektor enthält die 1000 höchsten Termgewichte. Über diese Repräsentation bietet sich die Möglichkeit an, ein Dokument mit seinen wichtigsten und spezifischen Wörtern zusammenzufassen. Da das TF/IDF Termgewicht den Fokus auf Wörter legt, die nicht in allen Dokumenten gleichermaßen vorkommen, geben die höchsten Termgewichte eines Dokuments die spezifischen Ter-

[7] Eine Zeitscheibe stellt in diesem Zusammenhang eine Menge von Dokumenten dar, die zu einer Sekunde, einer Stunde, einem Tag, einem Monat oder einem Jahr gehören.

me eines Dokuments wieder. An erster Stelle, mit den höchsten Gewichten, stehen meist Nomen oder Eigennamen, die Ereignisbezogen verwendet werden. In Tabelle 3.1 sind Ausschnitte einer Berechnung dargestellt. Die Tabelle zeigt, am Beispiel der Ereignisse von Fukushima im März 2011, Dokumente, die das Verfahren in einem Cluster zusammenfasst. Die Darstellung der Termvektoren kann einzeln für die Dokumente erfolgen, $\mathbf{w}_{d,t}$, oder als Zusammenfassung für alle Dokumente innerhalb eines Zeitintervalls. In der Tabelle sind jeweils die Zusammenfassungen eines Tages als gemittelter Termvektor, $\mathbf{w}_{avg,t}$, und als Termvektor mit den maximal vorzufindenden Gewichten unter allen Dokumenten eines Clusters, $\mathbf{w}_{max,t}$, dargestellt. Analog dazu werden für den gesamten Monat März die Mittelwert- und Maximalwert-Vektoren erstellt.

Die Darstellung in Tabelle 3.1 zeigt, dass die Gewichtung einen Fokus auf ereignisbezogene Nomen setzt. Dies entspricht der Erwartung, da das TF/IDF-Maß die Terme höher gewichtet, die in wenigen oder keinen anderen Dokumenten vorhanden sind. Alle anderen relevanten Wortarten, wie Verben oder Adjektive, können in der Berichterstattung über andere Ereignisse verwendet werden und sind demnach, im Gegensatz zu Nomen, weniger spezifisch für ein Ereignis oder eine Thematisierung. Somit ist die detaillierte Interpretation thematischer Inhalte nach diesem Verfahren im Wesentlichen auf Nomen beschränkt.

Die Bildung von Zusammenfassungen über mehrere Dokumente ist durch die Repräsentation als Vektoren möglich. Die Darstellung mit einem Vektor, der den Durchschnitt aller Vektoren eines Clusters repräsentiert, $\mathbf{w}_{avg,t}$, kann als Querschnitt verstanden werden. Terme, die in allen ereignisbezogenen Dokumenten manifestiert sind, werden auch in einem Durchschnittsvektor enthalten sein. Damit wird eine Sichtweise generiert, welche die gemeinsamen Verbindungen oder Referenzen eines Ereignisses, und somit eines Themas, repräsentiert. Dies entspricht den Termen mit wenig Dynamik, die ein Thema, also die mangelhaften Objekte, in einem Text darstellen. Die Darstellung der maximalen Gewichte in einem Vektor, $\mathbf{w}_{max,t}$, erlaubt dagegen eine Zusammenfassung, die spezifische, in einzelnen Dokumenten vorkommende Elemente enthält. So können Einzelaspekte eines Clusters dargestellt werden, die in wenigen Dokumenten einer Thematisierung angesprochen werden. In Tabelle 3.1 ist bei der Zusammenfassung der Maximalgewichte des gesamten Monats März (3. Zeile der Tabelle) ersichtlich, dass auch Bestandteile wie Kaiser Akihito (emporer, Akihito) und die Währung Yen als Einzelaspekte innerhalb der Thematisierung zu finden sind.

2011-03-11	$\mathbf{w}_{d,t}$	reactors:6,521, nuclear:6,402, reactor:6,324, cooling:5,557, plant:5,155, radiation:3,641, radioactive:2,968, Japanese:2,955, safety:2,764, earthquake:2,742, fuel:2,297, leak:2,182, evacuation:2,09, plants:1,983, meltdown:1,938, generator:1,92, systems:1,874, Tokyo Electric Power Company:1,85, Japan:1,8, diesel:1,689
	$\mathbf{w}_{avg,t}$	GMT:9,179, nuclear:6,963, reactor:6,957, plant:6,505, reactors:5,869, Japan:5,041, Tepco:4,994, cooling:4,075, tsunami:3,714, radioactive:3,561, radiation:3,034, Fukushima Daiichi:2,938, earthquake:2,895, Fukushima:2,632, Japanese:2,463, plants:2,38, quake:1,852, vapor:1,763, Kyodo:1,753
	$\mathbf{w}_{max,t}$	GMT:27,537, reactor:12,649, plant:11,782, Tepco:11,457, Japan:11,162, nuclear:10,782, tsunami:9,55, reactors:7,826, radioactive:7,122, Fukushima Daiichi:7,05, Fukushima:6,318, cooling:5,557, earthquake:5,485, radiation:4,855, Kyodo:4,507, Japanese:4,433, vapor:4,406, plants:4,363, agency:4,16
2011-03-15	$\mathbf{w}_{d,t}$	radiation:19,951, plant:15,331, Japan:10,819, nuclear:9,876, reactor:9,116, Tokyo:8,078, fuel:7,318, reactors:7,092, rods:6,952, Fukushima:6,528, Japanese:6,495, levels:6,02, radioactive:5,949, tsunami:5,941, unit:5,871, Kyodo:5,831, radius:5,691, Radiation:5,338, millisieverts:5,101, milliSieverts:5,20
	$\mathbf{w}_{avg,t}$	nuclear:3,536, radiation:2,415, reactor:2,065, reactors:1,991, plant:1,951, Japan:1,934, Fukushima:1,098, Japanese:1,094, fuel:1,082, rods:1,043, tsunami:0,983, Tokyo:0,83, earthquake:0,791, radioactive:0,757, Kyodo:0,745, power:0,709, disaster:0,665, fire:0,641, levels:0,635, Tepco:0,619
	$\mathbf{w}_{max,t}$	radiation:19,951, plant:15,331, nuclear:13,699, Japan:10,819, reactor:10,189, reactors:8,183, Tokyo:8,078, rods:7,532, fuel:7,318, Fukushima:6,528, Japanese:6,495, Fukushima Daiichi:6,478, levels:6,02, radioactive:5,949, tsunami:5,941, unit:5,871, Kyodo:5,831, radius:5,691, Radiation:5,338, earthquake:5,276
2011-03	$\mathbf{w}_{avg,t}$	nuclear:2,821, reactor:2,394, radiation:2,064, plant:1,958, Japan:1,88, reactors:1,608, tsunami:1,157, Fukushima:1,148, Tepco:1,022, Japanese:0,935, JST:0,859, radioactive:0,858, earthquake:0,816, fuel:0,703, water:0,682, rods:0,608, Tokyo:0,598, power:0,553, disaster:0,54, safety:0,519
	$\mathbf{w}_{max,t}$	JST:33,068, GMT:27,537, radiation:19,951, reactor:19,912, nuclear:18,774, Tepco:16,206, plant:15,331, Japan:15,083, Akihito:11,623, Japanese:11,124, tsunami:10,892, Fukushima:10,368, rods:9,566, emperor:9,528, reactors:9,46, Fukushima Daiichi:8,809, earthquake:8,719, explosion:8,557, prefecture:8,527, yen:8,213

Tabelle 3.1: Beispiele für TDT-Termvektoren, welche aus einem Ausschnitt der Online-Ausgabe des Guardian (März 2011) erstellt wurden. Das Korpus enthält 11.000 Dokumente mit insgesamt 5,2 Mio. Token. Die zwei oberen Darstellungen zeigen jeweils einen Tag der Berichterstattung. Die untere Darstellung bezieht sich auf den gesamten März 2011.

Demnach ergibt die Reflexion des Verfahrens bezüglich der linguistischer Theorien, dass die Möglichkeiten der Darstellung und Interpretation hauptsächlich auf Nominalgruppen, wie durch Fritz (1982) und Lötscher (1987) beschrieben, beschränkt sind. Die Sinnkomplexe oder die mangelhaften, zu erklärenden Objekte, stellen sich als sortierbare Nominalgruppe dar. Nach Fritz (1982) sollen die Nomen gleichberechtigt sein. Dieses Konzept wird durch die Darstellung der Vektoren und die Termgewichte erweitert. Durch die Gewichtung lässt sich beurteilen, welche Nomen bzw. Terme spezifischer für eine Thematisierung oder ein Ereignis sind. Dennoch kann die Darstellung nur eine Zusammenfassung der zu erklärenden Objekte liefern und lässt aus, wie die Objekte innerhalb des Themas erklärt und referenziert werden. Beispielsweise ergibt sich aus der Darstellung nicht, warum einerseits Brennstäbe (rods) und andererseits ein Erdbeben (earthquake) im Zusammenhang mit einem Kraftwerk (plant) genannt werden. Die Expansion oder Entfaltung des Themas, nämlich dass die Brennstäbe nicht gekühlt werden können und das Kraftwerk durch ein Erdbeben beschädigt wurde, können ohne zusätzliches Wissen nicht nachvollzogen werden. Die semantischen Kontexte der Nomen bleiben vorerst unklar und müssen anders erfasst werden. Es ist nicht möglich, Nebenthemen oder thematische Sprünge in den Texten zu identifizieren, da bei dem Verfahren nur ein Cluster für einen Text vergeben werden kann. Vorstellbar ist hier, dass die Clusterbildung auf Absätzen durchgeführt wird, welche die Analyseeinheit, anstelle der Dokumente, darstellen. Diese Idee stellt sich allerdings als problematisch dar. Es müssen sehr viele Dokumente, die aus den Absätzen gewissermaßen erzeugt werden, verarbeitet werden. Die Termvektoren, die für die Ähnlichkeitsberechnungen benötigt werden, basieren auf wenigen Termen eines Absatzes und sind dadurch sehr dünn besetzt. Die Kosinus-Ähnlichkeit und die TF/IDF-Termgewichtung sind kaum abhängig von der Dokumentlänge. Deshalb führen sehr kurze Texte, auch bei einer geringen Überdeckung von Vokabular, zu hohen Ähnlichkeiten bei der Kosinus-Ähnlichkeit. Dagegen haben lange Texte oder Absätze mit viel Vokabular wenig Ähnlichkeit zu kurzen Texten, obwohl sich, absolut gesehen, mehr Vokabular decken kann. Dies führt zu dem Effekt, dass gesetzte Schwellwerte nicht mehr zuverlässig arbeiten und sehr kurze Absätze meinst in einem Cluster gruppiert werden. Dagegen verbleiben lange Absätze oft einzeln in den resultierenden Clustern. Demnach kann nicht zuverlässig und unabhängig von der Länge der Absätze geclustert werden. Die Idee, dass dieses Verfahren auf sehr kurze Texte, wie Einzelabsätze angewendet werden kann, muss deshalb verworfen werden. Eine Veränderung des Schwellwertes, der beim Vergleich der Dokumente überschritten werden muss, sollte es erlauben gröbere Thematisierungen oder ereignisübergreifende

2011-03	$\mathbf{w}_{d,t}$	nuclear:7,412, power:2,821, coal:2,474, plants:2,335, plant:2,161, coal-burning:2
	$\mathbf{w}_{d,t}$	Tepco:15,571, reactor:12,579, plant:10,372, radiation:8,696, Fukushima Daiichi:8,435, nuclear:7
	$\mathbf{w}_{d,t}$	Christchurch:4,61, earthquake:3,073, city:2,591, quake:2,305, New Zealand:1,537, funeral:1,537
	$\mathbf{w}_{d,t}$	nuclear:2,726, stations:1,954, power:1,853, reactors:1,724, industry:1,72, waste:1,461, Chernobyl:1

Tabelle 3.2: Darstellung unterschiedlicher Dokumente eines Clusters. Durch den geringen Schwellwert von 0.1 werden auch Dokumente anderer Stories zugeordnet. So können abstraktere Themen gebildet werden. Bei einer solchen extremen Wahl werden unabhängige Themen miteinander vermischt und die erstellten Themen sind nicht interpretierbar.

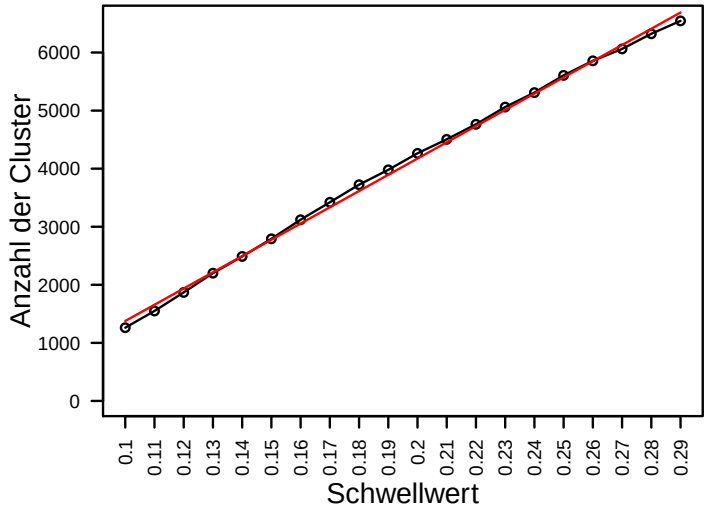

Abbildung 3.2: Darstellung der Anzahl der produzierten Cluster im Testkorpus, welche abhängig vom Schwellwert für die Übereinstimmung zweier Cluster gebildet wurden.

Dokumente zusammenzufassen. Das Verhalten des Verfahrens im Hinblick auf die Veränderung des Schwellwertes spiegelt dessen Fähigkeit wieder, die Granularität und Abstraktion der dargestellten Thematisierungen, durch eine Parametrisierung des Schwellwertes zu beeinflussen. Die Veränderung des Schwellwertes führt dazu, dass die Anzahl der resultierenden Cluster beeinflusst wird. Auch die durchschnittliche Anzahl der Token, die innerhalb eines Clusters liegen, verändert sich, da weniger oder mehr Dokumente zugeordnet werden. Die Veränderung der Menge der resultierenden Cluster und der Token innerhalb eines Clusters verhält sich linear zur Anpassung des Schwellwertes, wie in Abbildung 3.2 und Abbildung 3.3 abzulesen ist. Die lineare Veränderung zeigt, dass durch eine kontinuierliche Anpassung des Schwellwertes auch der Abstraktionsgrad innerhalb der Cluster und die Anzahl der Themen linear

angepasst werden kann. Denn die Anzahl der Token, die einem Cluster zugeordnet werden, entscheidet über dessen Darstellungsgranularität. Die Abstraktion und der Grad der Zusammenfassungen wird in den Tabellen 3.2 und 3.3 an zwei extremen Einstellungen für den Schwellwert deutlich. Bei einem Schwellwert von 0.1 ist auffällig, dass unterschiedliche Themen, die aber untereinander Ähnlichkeiten aufweisen zusammengefasst werden. So finden sich zusammen mit den Ereignissen in Japan auch Dokumente zu dem Erdbeben in ChristChurch, New Zealand, in diesem Cluster wieder. Es zeigt sich, dass der Begriff New Zealand zu Ähnlichkeiten mit der Sport-Berichterstattung (Rugby) führt. Dieser Effekt wird beseitigt, sofern zwei separate Schwellwerte für die Eigennamenvektoren und das Restvokabular verwendet werden (Allan u. a., 2005, vgl.). Diese können individuell angepasst werden. Dennoch werden bei einem zu geringen Schwellwert Dokumente assoziiert, die inhaltlich nichts miteinander zu tun haben und nur aufgrund einer geringen Übereinstimmung einander zugeordnet werden. Die Berechnung von $\mathbf{w}_{avg,t}$ ist bei diesen groben und großen Clustern nicht unproblematisch, da bei einem sehr großen Cluster, welches von einer Story dominiert wird, auch der Durchschnittsvektor verzerrt wird und nicht alle Aspekte eines Clusters gleichberechtigt wiedergegeben werden. Die Problematik eines sehr niedrigen Schwellwerts wird in Tabelle 3.2 genauer dargestellt. Wird der Schwellwert höher als die als Optimum vorgeschlagene Referenz von 0.21 gesetzt, so enthalten die Cluster im Durchschnitt weniger Dokumente und bilden nur Teilaspekte einer Story ab. Anhand des Beispiels in Tabelle 3.3 wird dargestellt, wie sich das Cluster für die Ereignisse in Fukushima bei einem Schwellwert von 0.29 noch einmal in zwei unterschiedliche Gruppen von Dokumenten aufteilt. Es entsteht ein Cluster für die Folgen der Naturkatastrophe und ein Cluster für die wirtschaftlichen Folgen. Durch die Erhöhung des Schwellwertes werden demnach wichtige Einzelaspekte einer Story oder eines Themas identifiziert. Aber auch hier besteht die Gefahr von Fehlern. So gehen bei einer zu großen Schwelle Zuordnungen verloren und relevante Dokumente werden nicht zugeordnet. Die Verwendung eines geeigneten Schwellwertes für die Analyse muss deshalb bei jeder Arbeit erprobt und getestet werden. Die Verwendung des Verfahrens von Allan u. a. (2005) ermöglicht, Dokumente zu Gruppen gleichartiger Stories zusammenzufassen. Im Folgenden wird die Menge D_k als Menge aller zu einer bestimmten Thematisierung gehörenden Dokumente verstanden. Die Menge aller Dokumente D_k, entspricht bei diesem Verfahren der Zuordnung D_C der Dokumente in einem Cluster. Durch die Darstellung der Dokumente als Vektoren ist es möglich, dass Themen- bzw. Story-Zusammenfassungen über Dokumentgrenzen hinweg in Form von Durschschnittsvektoren gebildet werden können. Dies lässt ei-

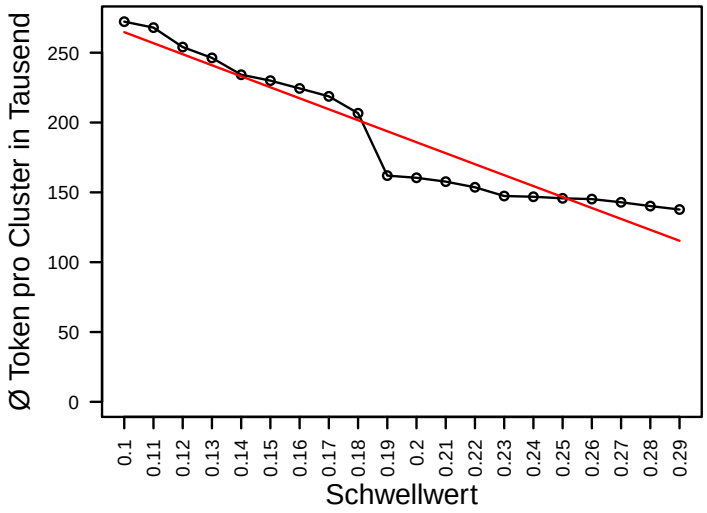

Abbildung 3.3: Durchschnittliche Anzahl der Token innerhalb eines Clusters, welche abhängig vom Schwellwert für die Übereinstimmung zweier Cluster gebildet wurden.

2011-03	$\mathbf{w}_{avg,t}$	yen:3,799, Japan:2,434, Japanese:2,213, markets:1,932, G7:1,724, earthquake:1,447
	$\mathbf{w}_{avg,t}$	nuclear:3,469, Japan:2,576, plant:2,348, reactor:2,192, radiation:1,95, tsunami:1,518

Tabelle 3.3: Durchschnittsvektoren zweier Cluster, die sich bei einem hohen Schwellwert von 0.29 aus der Fukushima Thematisierung separieren.

ne Darstellung von Themen als Wortgruppe oder Nominalgruppe zu. Um den aus einem Korpus extrahierten Themen einen gewissen Abstraktionsgrad vorzugeben, lässt sich der Schwellwert des Verfahrens anpassen. Der verwendete Schwellwert muss getestet werden, um falsche oder lückenhafte Assoziationen zu vermeiden. Durch die Verarbeitung kompletter Dokumente, ist eine Mehrfachzuordnung von Stories oder Themen zu einem Dokument, nicht möglich. Eine Beobachtung von sich bedingenden, assoziierten Themen oder Themensprüngen ist nicht möglich.

3.2 Topic-Modelle

Natürlichsprachige Texte werden auf Grundlage eines oder mehrerer thematischen Zusammenhänge verfasst. Diese Annahme wird durch die funktionale Satzperspektive, die Fokustheorie und die Abstraktion auf Makropropositionen formal abgebildet. Die kategoriale Abgrenzung und die Abbildung des Themas in Nominalgruppen bildet Themenwechsel innerhalb der Dokumente nicht explizit ab und es ist unklar, wie

Themenwechsel erfasst werden können. Einerseits können mehrere Nominalgruppen in Dokumenten erfasst werden. Andererseits ist aber unklar, wie diese Abgrenzung zu erreichen ist. Das TDT-Verfahren erzeugt lediglich eine Clusterzuordnung für ein analysiertes Dokument. Um mehrere Cluster für ein Dokument vergeben zu können, müssen die Dokumente separiert werden, wobei die Länge der abgetrennten Einheiten frei wählbar ist.

Das Thema, welches von einem Autor gewählt wird, bestimmt direkt, welche Worte für die Produktion eines Textes Verwendung finden. Somit ist ein Text, der zu einem bestimmten Thema verfasst wird, immer eine Mischung aus Worten, die für die Darstellung eine Themas nötig sind. Gewissermaßen eine Mischung aus den zu erklärenden Objekten und den Erklärungen selbst. Die Verwendung der Wörter teilt sich in eine syntaktische Wortgruppe und eine semantische Wortgruppe auf, wobei die syntaktischen Worte eine „[. . .]gedankliche Formung zum Ganzen durch Beziehungssetzung zwischen den Bestandstücken der Darstellung [. . .]“ (Bergenholtz u. Schaeder, 1977, vgl. S. 28) sind. Sie zählen „[. . .]zum objektiven Teil, zur Zusammenfassung und Organisation der Wirklichkeit, und werden als Ganzheitselemente des Satzes bezeichnet, z.B. ‘und, als, dem’.“ (Bergenholtz u. Schaeder, 1977, vgl. S. 28). Während die syntaktischen Wörter demnach eine grammatische Funktion erfüllen, repräsentieren die semantischen Worte die bedeutungstragenden Eigenschaften eines Textes und dienen damit als Stellvertreter für die thematischen Refrenzen oder mangelhaften, zu erklärenden Objekte im Text.

3.2.1 Latent Dirchlet Allocation

Der Annahme folgend, dass Texte auf mehreren Themen basieren können, wurden statistische Modelle entwickelt, die unter dem Begriff der Topic-Modelle zusammengefasst sind. Ein erstes Modell dieser Art wird als Probabilistic latent semantic indexing (pLSA) (Hofmann, 1999) bezeichnet. Dieses Modell erweiterte den Ansatz des Latent semantic indexing (LSI) (Deerwester, 1988; Deerwester u. a., 1990) um eine statistische Grundlage, welche die Probleme des LSI Verfahrens addressiert. Bei der Verwendung von LSI werden alle Dokumente in einem Korpus in das Termvektor-Modell überführt, sodass eine Dokument-Term-Matrix entsteht. Die Entstehung der Termvektoren kann auf der Grundlage unterschiedlicher Verfahren stattfinden. Einfache Termvektoren werden mit lokalen Termgewichtungen, die entweder auf der Termhäufigkeit im Dokument basieren oder auf einem Binärwert, der lediglich die Existenz eines Terms in einem Dokument anzeigt, gebildet. Die so generierte Term-Dokument-Matrix wird mittels Dimensionsresuktion in einen latenten semantischen Raum projiziert, sodass

die Varianz der originalen Matrix erhalten bleibt. Das LSI Verfahren und die Reduktion der Term-Dokument-Matrix fussen auf einer Least-Squares Methode (Manning, 1999, vgl. S. 558 f.). Im niedrig-dimensionalen Raum ist es möglich, die Dokumente in einem Korpus anhand ihrer Dokument-Kookkurrenzen zu vergleichen. Terme, die in verschieden Dokumenten gemeinsam auftreten, führen bei den entsprechenden Dokumenten zu ähnlichen Vektoren im latenten semantischen Raum. Diese Funktionalität bietet die Möglichkeit, Dokumente anhand der gemeinsamen Wortverwendung zu vergleichen. Hofmann u. a. (1999) stellt allerdings fest, dass die Darstellungen der Dokumente als Vektoren im latenten semantischen Raum, also dimensionsreduziert, nicht interpretiert werden können, sodass die Wortverwendungen, die zu einer Dokumentähnlichkeit führen, nicht analysiert werden können. Das Verfahren ist demnach für den Dokumentvergleich aber nicht für eine thematische Analyse der Dokumente nutzbar. Die Schätzung bzw. die Zerlegung der Dokument-Term-Matrix basiert beim LSI Verfahren auf einer Regression bzw. einer L2-Norm. Im Gegensatz dazu basiert das pLSI Verfahren auf der Wahrscheinlichkeitsfunktion eines multinomialen Samplings. Das bedeutet, dass alle Modellbestandteile definierte und normalisierte Wahrscheinlichkeiten besitzen. Während die reduzierte Darstellung des LSI Verfahrens keine Interpretation erlaubt, kann die reduzierte Darstellung des pLSI Modells als multinomiale Wortverteilung gesehen werden (Hofmann, 1999). Ein weiterer Vorteil dieser statistischen Darstellung ist eine optimale Selektion der Anzahl der Klassen bzw. der latenten semantischen Dimensionen, mit Hilfe statistischer Theorie, was hingegen beim LSI Verfahren intuitiv und empirisch bewerkstelligt werden muss. Um eine vollständige statistische Modellierung des Problems zu realisieren, nutzt das pLSI Verfahren ein so genanntes Aspect Model (Hofmann u. a., 1999). Diese Modelle assoziieren eine nicht beobachtbare latente Variable (latent) mit manifesten Variablen vorliegender Daten. Eine beobachtbare Größe kann in diesem Fall eine Wortmenge eines Dokuments sein. Die Wörter dieser Wortmenge gehen untereinander eine Beziehung ein, da sie in einem Dokument gemeinsam verwendet werden. Daraus lassen sich Dokument-Wort Paarungen erzeugen. Diese Paarungen können beobachtet werden und werden mit der latenten Variable oder Klasse assoziiert. Das Aspect Model ist ein zufälliger generativer Prozess, der die Entstehung von gemeinsam auftretenden Daten beschreibt. In Hofmann u. a. (1999) ist der Prozess definiert als:

1. Wähle einen Aspekt z mit der Wahrscheinlichkeit $p(z)$,
2. Selektiere ein X Objekt $x \in X$ mit der Wahrscheinlichkeit $p(x|z)$ und
3. Selektiere ein Y Objekt $y \in Y$ mit der Wahrscheinlichkeit $p(y|z)$.

2011-03-11	$p(\mathbf{w}\|z)$	Japan, tsunami, earthquake, nuclear, people, GMT, plant, Tokyo, quake, waves, emergency, Pacific, coast, Japanese, power, reactor, areas, buildings, Sendai, magnitude, news, reactors, hit, residents, cooling, damage, agency, coastal, told, government
2011-03-23	$p(\mathbf{w}\|z)$	Libya, Gaddafi, forces, military, air, Libyan, strikes, people, regime, government, US, Libyan, Muammar Gaddafi, president, Obama, Tripoli, country, Benghazi, no-fly, war, Misrata, mission, rebels, rebel, resolution, zone, told, civilians, force, city

Tabelle 3.4: Beispiele für die Darstellung der latenten Verteilung $p(\mathbf{w}|z)$ für die Berichterstattung der Online-Ausgabe des Guardian im März 2011.

Im Fall einer Dokumentkollektion kann X als Dokumentkollektion und Y als Vokabular interpretiert werden. Diese Variablen werden im weiteren Verlauf als D und W weitergeführt. Die Wahrscheinlichkeit eines gemeinsam auftretenden Paares von x und y wird nach diesem Modell als Mischung aus Multinomialverteilungen beschrieben.

$$p(w,d) = \sum_z p(z)p(d|z)p(w|z) = p(d)\sum_z p(z|d)p(w|z) \tag{3.4}$$

Alternativ kann diese Verteilung auch für diskrete Datenpunkte als $p(w_i|d_j) = \sum_{k=1}^{K} p(z_k|d_j)p(w_i|z_k)$ angegeben werden.

Die Wahrscheinlichkeiten dieser Mischung müssen für einen realen Textkorpus mittels eines EM-Algorithmus (Hofmann, 1999) geschätzt werden. Im Gegensatz zum LSA Verfahren, werden die Dokumente hier nicht durch einen Vektor in einem dimensionsreduzierten Raum dargestellt, sondern durch Multinomialverteilungen über das Vokabular. Diese Verteilungen sind einerseits für die Dokumentklassifikation einsetzbar und können, anders als bei LSA, direkt als Wortliste interpretiert werden, sodass neben der statistischen Modellierung auch eine Strukturanalyse einer Dokumentkollektion anhand der Verteilungen möglich ist (Hofmann, 1998). Die Verteilungen $p(d|z)$ und $p(w|z)$ sind dimensionsreduzierte Darstellungen einer Dokumentkollektion.

Aus Sicht der Inhaltsanalyse liefert pLSA das erste unüberwachte Verfahren, dass Strukturanalysen auf Grundlage von empirischen Verteilungen einer Dokumentkollektion liefern kann. Die Verteilungen $p(w|z)$ aus den Aspekten z stellen hier semantische Zusammenhänge von Wörtern dar. In diesem Sinne können, nach der Definition der kategorialen Abgrenzung, Aspekte als Thema interpretiert werden, da die Verteilung $p(w|z)$ einen semantischen Zusammenhang der Wörter herstellt. Die Beschränkung auf Nomen ist durch das Verfahren selbst nicht gegeben. In Tabelle 3.4 sind beispielhaft Wortlisten aus Modellverteilungen $p(w|z)$ dargestellt, die aus einer Analyse der Online-Ausgabe des Guardian entstanden sind. Die Darstellung basiert auf den

aufsteigend sortierten Wahrscheinlichkeiten der Komponenten einer multinomialen Verteilung und deren assoziierten Wörtern.

Das pLSA Modell arbeitet mit einer latenten Variablen z. Diese Variable erfasst das gemeinsame Auftreten von Dokumenten und Wörten in den Multinomialen $p(w|z)$, $p(d|z)$ oder $p(z|d)$. Die Verteilung $p(z|d)$ impliziert, dass ein Dokument in der Kollektion im pLSA Modell mehrere Themen enthalten kann. Diese Verteilung muss für alle Dokumente in der Dokumentkollektion errechnet werden, was zu Problemen führt, da das pLSA Modell kein „[...]wohl definiertes generatives Modell für Dokumente ist; es existiert kein natürlicher Weg, zuvor nicht trainierten Dokumenten eine Wahrscheinlichkeit zuzuordnen." (Blei u. a., 2003). Das bedeutet, dass mit pLSA zwar statistische Modelle und Strukturanalysen durchführt werden können. Die Möglichkeit, neue Dokumente anhand eines trainierten Modells, mit einer Wahrscheinlichkeit für die latente Variable z zu versorgen, fehlt allerdings. Blei u. a. (2003) verweisen auf ein Overfitting Problem, dass durch die zu inferierende Parametermenge entsteht. Die Verteilungen $p(w|z)$ und $p(d|z)$ sind jeweils von der Länge V (Vokabular) bzw. M (Dokumente) und müssen für jedes $z_k \in \{k = 1, \ldots, K\}$ (Komponenten, Themen) inferiert werden. Das heißt, dass $kV + kM$ Parameter geschätzt werden müssen und die Parameterlänge linear mit jedem neuen Dokument wächst. Um diese Probleme zu umgehen und um ein wohl definiertes generatives Modell zu definieren, wurde das Latent Dirichlet Allocation (LDA) Modell entwickelt (Blei u. a., 2003). Diese Arbeit stellt den Anfang einer Reihe von Arbeiten und Forschungen mit Topic-Modellen dar, welche die Schwächen von pLSA kompensieren. Das Modell definiert einen generativen Prozess für Dokumente.

1. Ziehe K Multinomiale $\phi_k \propto Dir(\beta_k)$, eines für jedes Thema k.
2. Für jedes Dokument d, $d = 1, \ldots, D$,
   (a) ziehe eine Multinomialverteilung $\theta_d \propto Dir(\alpha_d)$.
   (b) Für jedes Wort w_{dn} in einem Dokument d, $n = 1, \ldots, N_d$
      i. ziehe ein Thema $z_{dn} \propto Multinomial(\theta_d)$,
      ii. ziehe ein Wort w_{dn} aus $p(w_{dn}|\phi_{z_{dn}})$, die Multinomialverteilung bedingt von Thema z_{dn}.

Die Parameter β und α sind hier die sogenannten Hyperparameter des Modells, da sie die Grundvorraussetzungen des Modells beeinflussen. Dem generativen Prozess

folgend, können die latenten Variablen für jedes Dokument über eine a posteriori-Verteilung gefunden werden.

$$p(\theta, \phi, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \phi, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{3.5}$$

Der wesentliche Vorteil dieses Modells ist, dass nicht jedem Dokument eine Verteilung der Themen-Mischung $p(z|d)$ zugeordnet wird, sondern eine verborgene Zufallsvariable in den Dokumenten eingeführt wird. Die Beschaffenheit dieser Zufallsvariable wird über ein Sampling aus einer Dirichlet-Verteilung bestimmt, sodass für jedes Thema K eine Wahrscheinlichkeit für den gesamten Korpus geschlossen wird. Der Parameter θ entspricht einer Multinomialverteilung, da die Dirichlet-Verteilung die konjugierte A-priori-Verteilung eben dieser ist. Diese Erweiterung erlaubt es, die verborgene Zufallsvariable und deren Verteilung auf neue Dokumente anzuwenden, was dieses Modell generell anwendbar auf unbekannte Dokumente macht. Des Weiteren wachsen die zu schätzenden Parameter nicht linear mit der Anzahl der Dokumente, sondern nur mit dem Vokabular. Die Parameteranzahl entspricht $k + kV$. Diese wichtige Erweiterung sorgt dafür, dass die Arbeit mit hierarchischen Bayes'schen Modellen dieser Art vermehrt eingesetzt wird.

3.2.2 Erweiterungen und alternative Modelle

Die Latent Dirichlet Allocation fokussiert sich auf die Modellierung der Inhalte von Texten. Das Ziel der Inferenz ist es, die Verteilungen der Themen und Wörter als Parameter des Modells zu schätzen. Die Anzahl der Themenverteilungen wird durch einen weiteren Parameter festgelegt. Die gesamte Textkollektion, die in das Modell einfließt, behandelt jedes Dokument gleichwertig und die Dokumente werden als unabhängige Datenpunkte betrachtet. Das führt dazu, dass unterschiedlich Aspekte von Textdokumenten ausgeblendet werden, die aber für eine Anwendung in der Themenanalyse Bedeutung haben können.

1. Die Anzahl der Themen K, die in ihrer Wortzusammensetzung geschätzt werden sollen, ist abhängig vom Korpus und von der Beschaffenheit der Texte. Wenn keine Vorstellungen über die Anzahl oder Zusammensetzung der Themen vorhanden ist, ist es schwer diese Auswahl zu schätzen.

2. Dokumente haben oft andere Abhängigkeiten. Nur Inhalte zu verwenden, um Zusammenhänge darzustellen, blendet die Autorenschaft, räumliche oder zeitliche Abhängigkeiten der Dokumente aus.

3. Die einzelnen Themen werden unabhängig voneinander behandelt, sodass Abhängigkeiten oder Hierarchien in den Themen nicht nachvollzogen werden können.

Die genannten Aspekte, die neben den reinen Inhalten eine Rolle bei der Produktion von Text und der darin enthaltenen Themen spielen, können für die Modellierung von Topic-Modellen in Betracht gezogen werden, um deren Qualität und Darstellung zu verbessern. In diesem Abschnitt sollen die Modelle kurz vorgestellt werden, sodass die Einordnung und der Nutzen unterschiedlicher Modelle nachvollziehbar ist.

Da bei unbekannten Daten nicht klar ist, wie spezifisch die Themen sind, muss dies erprobt werden und die optimale Anzahl der Themen für eine inhaltsanalytische Fragestellung gefunden werden. Für die Festlegung einer optimalen Aufteilung des Korpus in K Themen, müssen im LDA-Modell mehrere Parameter festgelegt werden. Die Wahl der Parameter α, β und K stehen immer in einem Zusammenhang. Da, wie weiter unten gezeigt, α und β jeweils festlegen, wie spezifisch die Verteilungen $p(z)$ und $p(\mathbf{w}|z)$ sind, muss der Parameter K daraufhin optimiert werden, sodass das Modell eine hohe Wahrscheinlichkeit in Abhängigkeit von den Daten erhält. Soll der Parameter K fixiert werden, so müssen wiederum α und β optimiert werden. Um den Parameter K nicht in Abhängigkeit der Hyperparameter der Dirichlet Funktionen festlegen zu müssen, wird ein Topic-Modell vorgeschlagen, welches mit dem hierarchischen Dirichlet Prozess arbeitet. Dieser stochastische Prozess wird verwendet, wenn unterschiedliche Daten oder Dokumente mit einer diskreten Variable unbekannter Kardinalität beschrieben werden müssen (Teh, 2006; Teh u. Jordan, 2010). Dies bedeutet, dass das Modell in der Lage ist, jedes Dokument mit einer anderen Dirichlet Verteilung, über die darin enthaltenen Themen, zu beschreiben. Diese Themen sind dennoch global indexiert und können über unterschiedliche Dokumente hinweg geteilt werden. Das HDP-LDA Modell legt die optimale Anzahl der Themen pro Dokument fest und kann damit die optimale Anzahl der Themen im gesamten Korpus festlegen. Die Aufteilung des Korpus in unterschiedliche Themen ist nur abhängig von den Parametern der Dirichlet Verteilungen im stochastischen Prozess. Auch in der HDP-LDA steuern die Parameter der Dirichlet Verteilungen die Konzentration der Wahrscheinlichkeitsmasse in den Multinomialen $p(\mathbf{w}|z)$. Je spezifischer sich die Themen auf ein geringes Vokabular konzentrieren sollen, desto mehr Themen müssen innerhalb des Korpus unterteilt werden. Somit ist in diesem Modell nur die Granularität der Themen festzulegen und das Modell bestimmt die Anzahl der Themen selbst. Das Modell besitzt, in Abhängigkeit von den Hyperparametern, immer die optimale Aufteilung in Themen.

Für die Modellierung der Dokumente einer Textkollektion, kann es hilfreich sein, andere Informationen aus den Metadaten der Dokumente in das Modell zu integrieren. Diverse Ansätze erlauben es, die Modellqualität, in Bezug auf unterschiedliche Aufgaben, zu verbessern. Unter anderem existieren Modelle, die Autorenschaft (Rosen-Zvi u. a., 2010), Ortsbezüge (Wang u. a., 2007) und zeitliche Abhängigkeiten der Dokumente (Wang u. McCallum, 2006; Masada u. a., 2009; Blei u. Lafferty, 2006) modellieren. Modelle, welche die Zeitstempel eines Dokuments modellieren, sind für die thematische Untersuchung von Nachrichtenkorpora interessant. Die Zeitstempel der Dokumente werden auf unterschiedliche Weise in die Modelle integriert. In der Arbeit von Wang u. McCallum (2006), wird das generative LDA Modell erweitert. Für jede Stelle n, in einem Dokument d, wird ein Thema $z_{dn} \propto Multinomial(\theta_d)$ und ein Wort w_{dn} aus $p(w_{dn}|\phi_{z_{dn}})$ ermittelt. In diesem Modell generiert das Thema z_{dn} einen Zeitstempel $t_{dn}|\psi_{z_dn} \propto Beta(\psi_z dn)$, der aus einer Beta-Verteilung ermittelt wird. In der Inferenz dieses Modells werden die Themen so über die Dokumente verteilt, dass der Anteil der Dokumente eines Themas im zeitlichen Längsschnitt Beta-verteilt ist. Durch die Eigenschaften der Beta-Verteilung kann die zeitliche Verteilung der Dokumente monoton-steigend oder -fallend sein oder ein globales Maximum aufweisen. Hat die Dokumentverteilung der Daten im Längsschnitt mehrere lokale Maxima, so stößt dieses Modell an seine Grenzen. Es ist nur für zeitlich begrenzte oder einem Trend unterliegende Themen geeignet. Der Ansatz von Masada u. a. (2009) fügt bei der Vorverarbeitung der Daten ein Array aus Zeitstempeln zum Vokabular der Dokumente hinzu. Dabei wird jeweils der Zeitstempel des verarbeiteten Dokuments genutzt. Diesen Zeitstempeln wird in der Inferenz ein Thema zugewiesen, was über ein eigenes Inferenzschema realisiert wird. Damit schaffen es die Autoren, Abhängigkeiten zwischen Wortverwendungen und Zeitstempeln in das Modell zu implementieren. Der Einfluss der Zeitstempel kann variiert werden, indem die Menge des Zeitstempel-Vokabulars bestimmt wird. Je mehr Einfluss die Zeitangaben der Dokumente bekommen sollen, desto größer wird die Anzahl der hinzugefügten Zeitstempel gewählt. Neben diesen zwei Modellen verfolgt der Ansatz von Blei u. Lafferty (2006) eine andere Strategie, die zeitlichen Abhängigkeiten in das Modell zu integrieren. Ausgehend von der Annahme, dass sich Themen und deren Inhalte über die Zeit verändern, wird dem Modell ermöglicht, die Themenverteilungen $p(\mathbf{w}|z)$ dynamisch an eine Zeitscheibe, wie z.B. die Dokumente innerhalb eines Tages, anzupassen. Dies geschieht im Modell über eine mögliche Veränderung der Wortwahrscheinlichkeit $p(\mathbf{w}|z)$ innerhalb der Themen einer Zeitscheibe. Innerhalb des LDA Modells wird hingegen angenommen, dass die globalen Themen jedem Wort jederzeit gleiche Wahrscheinlichkeit zuord-

nen, selbst wenn der Korpus diachron ist. Diese Dynamik der Wahrscheinlichkeiten wird modelliert, indem die Dynamik für jedes Wort, in einer Themenverteilung einer Normalverteilung unterliegt. Ferner wird in diesen Dynamic-Topic-Modellen (DTM) keine Dirichlet Verteilung genutzt, um daraus Multinomiale zu erzeugen, sondern die Logistische Normalverteilung (Aitchison u. Shen, 1980). Im DTM wird modelliert, dass sich der Mittelwert für die Wörter in den Themen in aufeinanderfolgenden Zeitscheiben verändern kann, sodass sich die Zusammensetzung der Themenverteilungen ändern kann. Eine sprachliche Dynamik kann abgebildet werden und die Themen müssen in diachronen Korpora nicht global fixiert sein. Blei u. Lafferty (2006) schlagen vor, dass die globalen Themenanteile im Korpus $p\mathbf{z_t}|M_t$ einer normalverteilten Dynamik unterliegen. Die Themenanteile können damit über die Zeit normalverteilt variieren. Das Modell passt sich besser an die Dynamik der Textdaten an und die Themenzuordnungen sind genauer, was sich auf die Qualität der Dokumentzuordnungen zu den Themen auswirkt. Im Gegensatz zu den anderen Modellen, die die Zeit berücksichtigen, wird in diesem Modell eben nicht nach einem bestimmten Verlauf der Themenproportionen oder Anteile in den Zeitschieben eines diachronen Korpus modelliert, sondern nach der inhaltlichen Dynamik. Bei der bloßen Modellierung der möglichen Themenanteile über die Zeit, bleiben die Themen dennoch inhaltlich fixiert und der wesentliche Aspekt der Veränderung wird außen vor gelassen.

In Textdokumenten werden verschiedene Themen gemeinsam verwendet. In den meisten Texten ist die Kombination unterschiedlicher Themen nicht zufällig und die Themen bedingen einander. Wird von einem Autor ein Thema gewählt, steigt die Wahrscheinlichkeit, dass bestimmte andere Themen ausgewählt werden. Soll in einem Nachrichtenartikel über die Folgen eines Erdbebens gesprochen werden, so wird über Erdbeben im Allgemeinen, Rettungsmaßnahmen, Politik, Folgen für die Wirtschaft oder andere Naturkatastrophen berichtet. Es existieren demnach Beziehungen und Abhängigkeiten zwischen gemeinsam verwendeten Themen. Solche Abhängigkeiten werden mit Topic-Modell-Ansätzen modelliert(Blei u. a., 2004; Blei u. Lafferty, 2007). Die Ansätze unterscheiden sich in ein hierarchisches Topic-Modell und ein korreliertes Topic-Modell (Blei u. Lafferty, 2007). Das hierarchische Topic-Modell modelliert die Abhängigkeiten der in Dokumenten gemeinsam verwendeten Themen durch einen Baum mit L Ebenen, bei dem jeder Knoten einer Themenverteilung entspricht. Der generative Prozess eines Dokuments wählt in diesem Modell zunächst einen Pfad in diesem Baum, sodass eine Menge aus L Themen entsteht, die Themenanteile θ am Dokument definieren. Entscheidend ist, dass die Themen und die Vorstellungen über deren syntaktische und semantische Allgemeinheit und Spezifizität hierarchisch

abgebildet wird (Blei u. a., 2004). Insbesondere wird modelliert, dass die Granularität der Themen von der Wurzel des hierarchischen Baums immer mehr abnimmt und die Themen in den Ästen des Baums immer spezieller werden. So enthält das Thema in der Wurzel immer die syntaktischen und funktionalen Eigenschaften der Dokumente, wie beispielsweise Stopwörter oder domänenabhängige Wörter, die in allen Dokumenten zu finden sind. Ausgehend davon verzweigen sich die Themen, die aufgrund dieses funktionalen Themas semantische Inhalte beschreiben. Dagegen modellieren die korrelierten Topic-Modelle einen anderen Zusammenhang gleichzeitig verwendeter Dokumente. Bei der hierarchischen Strukturierung der Texte wird immer davon ausgegangen, dass es ein abstraktes syntaktisches Thema gibt, welches wenig semantische Bedeutung reflektiert. Die korrelierten Topic-Modelle modellieren allerdings nicht, welche Themen gemeinsam in einer Hierarchie in den Dokumenten verwendet werden, sondern welche Themen untereinander, und vor allem mit welcher Kontingenz, in Zusammenhang stehen. Da das syntaktische Thema in allen Dokumenten zu finden ist, ist die Korrelation zu anderen Themen geringer, als beispielsweise der Zusammenhang zwischen semantischen Themen, die auf Grundlage eines Ereignisses gemeinsam in einem Dokument verwendet werden. Die Korrelation unter den Topics wird durch die Verwendung der Logistischen Normalverteilung realisiert. Sie weist, als a priori-Verteilung für Multinomiale, eine Kovarianzfunktion für alle Komponenten auf (Aitchison u. Shen, 1980; Blei u. Lafferty, 2007). Dadurch kann eine Inferenz des Modells und der Kovarianzmatrix der Logistischen Normalverteilung auf einem Textkorpus zeigen, welche Themen sich untereinander bedingen. So kann nicht nur festgestellt werden, in welcher hierarchischen Abstraktionsstufe Themen miteinander genutzt werden. Vielmehr kann analysiert werden, welche semantisch relevanten und inhaltstragenden Themen häufig gemeinsam genutzt werden.

Die Modellierung von Textdokumenten ist meist eine vereinfachte Darstellung der Dokumente. Die Themen repräsentieren die möglichen Verteilungen von Wörtern in den Dokumenten. Die Themenverteilung innerhalb der Dokumente stellt die Inhalte eines Dokuments vereinfacht dar. Das ausführlich vorgestellte LDA Modell modelliert diese Aspekte und vernachlässigt einige Dokumenteigenschaften. Auf Grundlage der Idee, die hinter der LDA steckt, wurden weitere Modelle entwickelt, die einzelne Dokumenteigenschaften mit modellieren und damit spezielle Anwendungen erlauben. Um die Verständlichkeit der weiteren Analysen und Anwendungen zu gewährleisten, konzentriert sich die weitere Arbeit allerdings auf die Anwendung und Untersuchung der Modelle LDA und HDP-LDA. Die Komplexität der Modelle ist überschaubar und es ist einfacher, die Inferenz und die Ergebnismengen zu analysieren und die

Anwendungen zu erläutern. Die Inferenzen sind effektiver zu berechnen, die Rechenzeit ist kürzer und eine höhere Praktikabilität der Verfahren ist damit gegeben.

3.2.3 Berechnung und Inferenz

Für Topic-Modelle gilt, dass die a posteriori-Verteilungen der latenten Variablen nicht analytisch berechnet werden können (Blei u. a., 2003; Griffiths u. Steyvers, 2004). Im Fall der LDA muss die a posteriori-Verteilung nach Formel 3.5 geschätzt werden. In den Publikationen unterschiedlichster Topic Modelle ist es üblich, dass die Schätzverfahren für die vorgeschlagenen Modelle definiert werden. Andere Arbeiten beschäftigen sich ausschließlich mit der Entwicklung effizienter Inferenzen für bereits publizierte Modelle. Für die Schätzung von Erwartungswerten und a posteriori-Verteilungen in Bayes'scher Statistik existieren unterschiedliche Verfahren, die sich in der Berechnung und Effizienz unterscheiden. An dieser Stelle sollen die Schätzverfahren und deren Effizienz vorgestellt werden, die sich für Topic-Modelle bewährt haben. Das verwendete Schätzverfahren hat Einfluss auf die Ergebnisdarstellung und die Dokumentmengen, die in akzeptabler Zeit berechnet werden können.

Die Schätzmethoden, welche in Topic-Modellen verwendet werden sind

- Markov Chain Monte Carlo (MCMC) und
- Variationelle Inferenz (VB).[8]

In der Veröffentlichung zur LDA von Blei u. a. (2003), wird ein Expectation-Maximization-Algorithmus und ein Variational-Inference-Algorithmus für die Inferenz entwickelt. Bei der Inferenz mit VB-Methoden, wird eine Familie von Verteilungen Q über die Modellparameter und latenten Variablen vorgeschlagen, um die wahre Verteilung P der a-posteriori-Verteilung zu approximieren (Bishop, 2006, vgl. S. 463). Die wahre Verteilung in diesem Mechanismus kann beschrieben werden durch

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + KL(q \parallel p). \tag{3.6}$$

Dabei beschreibt $\mathcal{L}(q)$ eine untere Schranke für die Wahrscheinlichkeit der Verteilung q in dem Modell. Die Kullback-Leibler Distanz (KL) zwischen der wahren Verteilung p und der vorgeschlagenen Verteilung q muss addiert werden, um die wahre Verteilung zu erhalten. Das Ziel von VB ist demnach, dass $\mathcal{L}(q)$ maximiert wird oder $KL(q \parallel p)$ minimiert wird. Die Inferenz für den LDA Algorithmus geht nach diesem Muster

[8] Diese Art der Inferenz kann als Variational Bayes bezeichnet werden.

vor und definiert variationelle Parameter, um die wahren Parameter des Modells zu approximieren. Die Aktualisierungen der variationellen Parameter basiert auf einer Ableitung der KL und folgt demnach dem Gradienten zur Minimierung dieser Funktion. Mit den aktualisierten Parametern, schätzt die Inferenz die Erwartungswerte der Daten neu, sodass die Parameter erneut optimiert werden können. Dieser Mechanismus wird iterativ betrieben, bis die Wahrscheinlichkeit des Modells im Hinblick auf die variationellen Parameter konvergiert. Die Methode stellt eine analytische Parameteroptimierung dar. Auch für HDP-LDA (Teh u. a., 2007), CTM und DTM werden Inferenzen mit der VB Methode vorgeschlagen.

In ihrer Arbeit über eine alternative Inferenz für das LDA Modell kritisieren Griffiths u. Steyvers (2004), dass die Aktualisierungen für die variationellen Parameter auf Digamma-Funktionen beruhen, die sehr rechenintensiv sind (Asuncion u. a., 2009). Als Antwort auf dieses Problem wird eine MCMC Methode in Form eines Gibbs-Samplers (GS) vorgeschlagen (Griffiths u. Steyvers, 2004; Porteous u. a., 2008). Gibbs-Sampler approximieren die a-posteriori-Verteilung, indem iterativ neue Werte z_i in einer Verteilung $p(\mathbf{z})$ aus einer Verteilung $p(z_i \| \mathbf{z}_{\backslash \mathbf{i}})$ ersetzt werden (Bishop, 2006, vgl. S. 542). Die Werte einer Verteilung z_i werden quasi unter Berücksichtigung aller anderen Komponenten der Verteilung aktualisiert. Innerhalb der beschrieben Inferenz für das LDA Verfahren, ersetzt der Gibbs-Sampler in einer Iteration alle Zuordnungen eines Themas, abhängig von allen anderen Zuordnungen in allen Dokumenten. Die Modellparameter können anhand der Themenzuordnung berechnet werden. Ein Gibbs-Sampler konvergiert durch stochastisches Ermitteln von Zuständen zur wahren Verteilung in einem Modell. Gibbs-Sampler werden über mehrere Iterationen hinweg initialisiert, dem sogenannten Burn-In, sodass die Verteilungen konvergieren können. Aus dem initialisierten Sampler können Samples erzeugt werden, aus denen die Modellparameter errechnet werden können. Für weitere Größenbestimmungen in Topic-Modellen kann es sinnvoll sein, die Themenzuordnungen einzelner Wörter vorliegen zu haben. Die Benutzung von Gibbs-Samplern ist deshalb für manche Anwendung sinnvoll, bei denen die Themenzuordnungen zu den Einzelwörtern eine Rolle spielt.

Die VB Inferenz für die LDA leidet unter der rechenintensiven Nutzung von Digamma-Funktionen. Teh u. a. (2006) schlagen deshalb einen kollabierten VB (CVB) Ansatz vor, bei dem die Aktualisierungen nicht mehr auf Digamma-Funktionen beruhen, da die Modellparameter aus der Inferenzrechnung marginalisiert werden. Im vorgeschlagenen Gibbs-Sampler von Griffiths u. Steyvers (2004) wird diese Vereinfachung angewandt. Die Erweiterung stellt eine Verbesserung dar, da die Berechnung

effizienter, die Modellqualität höher ist und die Inferenz schneller konvergiert. In Asuncion u. a. (2009) werden die vorgeschlagenen Inferenzalgorithmen gegenübergestellt und es kann gesagt werden, dass

- die Inferenz mit GS zu einer besseren Approximation der a-posteriori-Verteilung führt und weniger rechenintensiv als VB ist,
- durch die stochastische Natur des GS dennoch viele Iterationen nötig sein können, bis die Konvergenz erreicht ist,
- die Marginalisierung des GS auch im VB-Algorithmus eingesetzt werden kann, und so die analytischen Vorteile des VB zu effizienten, deterministischen und nicht stochastischen Ergebnissen führen.

Für sehr große Datenmengen skalieren die bisher besprochenen Inferenzmechanismen nicht ausreichend, da für eine Iteration oder eine Parameter-Aktualisierung alle Datensätze in die Berechnung mit einbezogen werden müssen. Trotz einer linearen Skalierung, genügen die Inferenzen den Anforderungen großer Datenmengen nicht ausreichend. Im Gibbs-Sampler muss jede Themenzuordnung für ein Wort aus K Themen ermittelt werden. Die $K \times N$ Berechnungen müssen für jedes Dokument in der Kollektion gemacht werden, um eine Iteration abzuschließen. Verändert sich die Anzahl der Dokumente oder der Themen im Modell um einen großen Faktor, so ist damit eine signifikant höhere Rechenzeit verbunden. In Abbildung 3.4 und Abbildung 3.5 ist dargestellt, wie sich das Laufzeitverhalten von Gibbs-Samplern verhält, wenn die Anzahl der Dokumente oder der Parameter K im LDA-Modell verändert werden. Die Berechnungszeit wächst mit Veränderung der Parameter linear. Ein ähnliches Bild ergibt sich bei VB Inferenzen, da dort alle Dokumente einer Kollektion in die Aktualisierung der Parameter in einer Iteration einbezogen werden. Mit einer als Stochastic Variational Inference (SVB) bezeichneten Methode, werden neue Inferenzverfahren für Topic-Modelle entwickelt, um bei der Komplexität der Inferenz nicht von der Datenmenge abhängig zu sein (Hoffman u. a., 2010; Wang u. a., 2011; Hoffman u. a., 2013). Die Parameteroptimierung der variationellen Parameter wird nicht durch eine Iteration aller Dokumente optimiert, sondern es wird sequenziell eine „lokale“ Aktualisierung auf Grundlage einer kleinen Untermenge von Dokumenten berechnet. Die Menge der Dokumente wird in kleine Untermengen aufgeteilt und die Parameter werden durch Gradienten, die aus den Untermengen gebildet werden, optimiert (Hoffman u. a., 2013). Die Inferenz betrachtet den Datensatz nur einfach, anstatt die gesamte Dokumentmenge in mehreren Iterationen zu beachten. Die Inferenz ist so wesentlich schneller und skaliert auf sehr großen Datensätzen besser. Andere

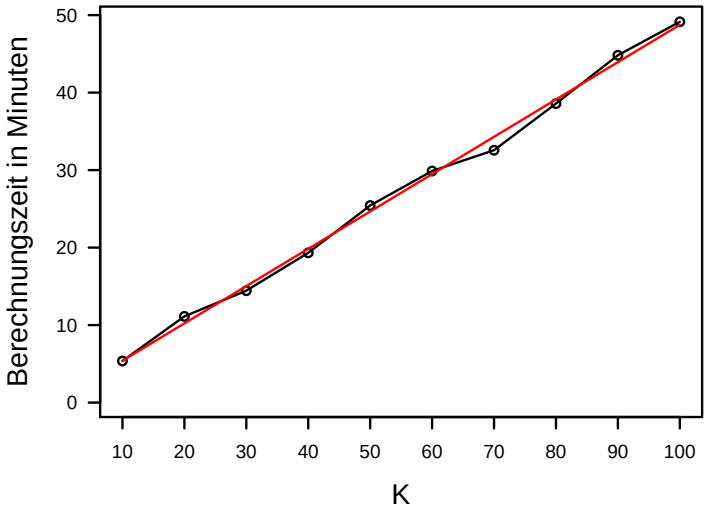

Abbildung 3.4: Verbrauchte Rechenzeit bei einem Gibbs Sampler für das LDA-Modell, bei 300 Dokumenten und einer Veränderung des Parameters K.

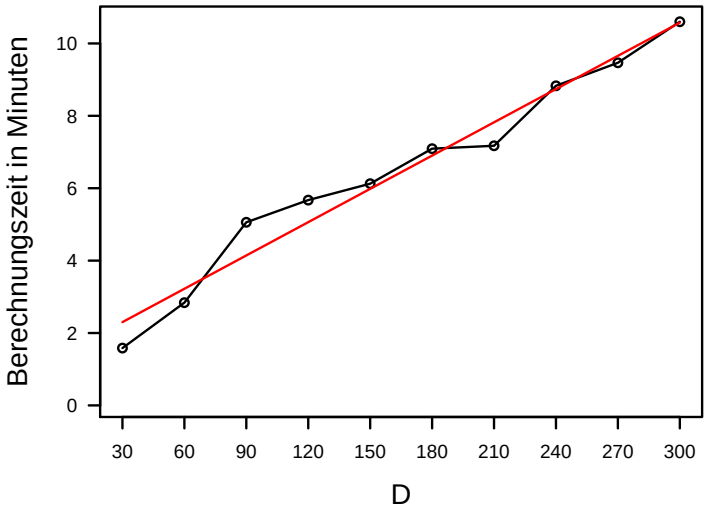

Abbildung 3.5: Verbrauchte Rechenzeit bei einem Gibbs Sampler für das LDA-Modell, bei unterschiedlichen Dokumentmengen und einem stabilen Parameter K.

Strategien, die Inferenzen skalierbar für große Datensätze zu machen, umfassen meist eine Parallelisierung der Berechnungen (Newman u. a., 2007, 2009; Wang u. a., 2009; Asuncion u. a., 2009; Yan u. a., 2009).

Die Wahl einer Inferenz hängt von der gewünschten Form des Ergebnisses und von der Menge der Daten ab. Während Gibbs-Sampler direkt mit den marginalisierten Verteilungen arbeiten und so zu einer gesuchten Verteilung konvergieren, optimieren die VB Methoden die Modellparameter direkt. Die Modellparameter müssen bei GS Methoden nachträglich geschätzt werden, während die Themen für die Einzelwörter in den Modellen direkt vergeben sind. Bei den VB Methoden ist es genau anders herum, sodass die Themenzustände der Wörter erst mit Hilfe der Modellparameter bestimmt werden müssen. Sofern nicht nur an den Modellparametern, sondern auch an der Struktur der Dokumente Interesse besteht, muss nach der Inferenz in einem weiteren Schritt geschätzt werden, welches Thema zu einem Wort zugeordnet werden kann. Dafür müssen sämtliche Themenzuordnungen jedes Token in allen Dokumenten der Kollektion, auch wenn diese sehr groß ist, noch einmal in $D \times N$ Schritten aus der Themenverteilung für das Wort geschätzt werden. Werden innerhalb der VB und SVB Methoden Verteilungen eingesetzt, bei denen , wie bei der CVB Methode, die Modellparameter marginalisiert sind kann die Modellqualität verbessert werden. Dies gilt auch für stochastische VB Methoden (Teh u. a., 2006, 2007; Sato u. a., 2012; Foulds u. a., 2013).

Wie viele Themen hat ein Korpus? - Exkurs zur Dirichlet-Verteilung und -Sampling und deren Bedeutung für die latenten Variablen im LDA Modell

Die Dirichlet Verteilung ist die konjugierte a priori-Verteilung der Multinomialverteilung auf der Simplex $S_k = (x_1, \ldots, x_K) : x_1 \geq 0, \ldots, x_K \geq 0, x_1 + \ldots + x_K = 1$ und ist eine Erweiterung der Beta-Verteilung auf den multivariaten Fall (Bishop, 2006, vgl. S. 76). Sie stellt eine Familie von multivariaten Wahrscheinlichkeitsverteilungen dar und kann als Verteilung über Multinomialverteilungen verstanden werden. Die Dirichlet Verteilung wird oft als $Dir(\alpha)$ ausgeschrieben, wobei α einen Vektor der Länge K darstellt und $\alpha_1 > 0, \ldots, \alpha_K > 0$ sein muss. Für die Anwendung im LDA Modell bedeutet dies, dass die Eigenschaften der multinomialen Wahrscheinlichkeitsverteilungen $p(\mathbf{w}|z)$ und $p(z|\theta)$ mit den Hyperparametern *alpha* und *beta* beeinflusst werden können. Somit werden die Eigenschaften des gesamten Modells beeinflusst, wobei die Parameter durchaus uniform sein können (Blei u. a., 2003). Die Dichtefunktion und die Eigenschaften der multinomialen Verteilungen, die aus der Dirichlet Verteilung erzeugt werden können, lassen sich gut darstellen, indem

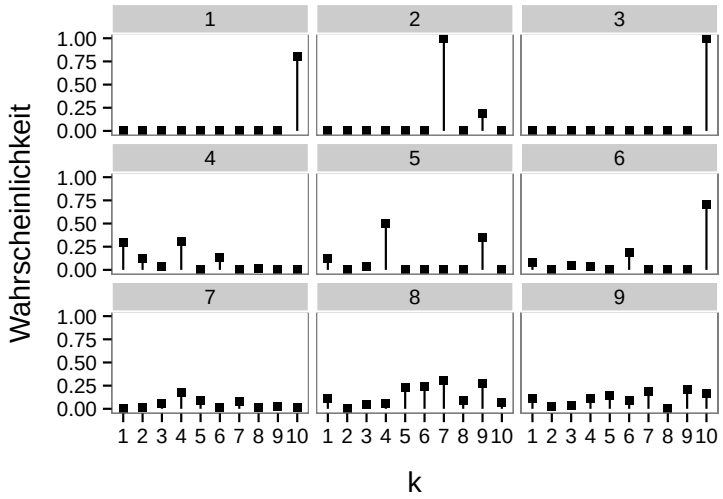

Abbildung 3.6: 9 Beispiele für Multinomialverteilungen, die aus einer Dirichlet-Verteilung ermittelt wurden. In den Reihen wurden jeweils die Werte für α auf Reihe 1: 0.01, Reihe 2: 0.2 und Reihe 3: 1 gesetzt.

beispielhaft verschiedene Verteilungen für unterschiedliche Werte von α erstellt und dargestellt werden. In Abbildung 3.6 sind 9 verschiedene Verteilungen dargestellt. In jeder Reihe sind 3 Beispiele für einen Wert von α dargestellt. In Reihe 1 wurde ein Wert von 0,01 gewählt, in Reihe 2 ein Wert von 0,2 und in Reihe 3 ein Wert von 1. An der Darstellung ist gut zu erkennen, dass bei einem Wert des Parameters $\alpha = 1$ Verteilungen erzeugt werden können, die allen 10 Klassen ähnlich viel Wahrscheinlichkeitsmasse zuordnen. Im Gegensatz dazu, kann am Fall $\alpha << 1$ festgestellt werden, dass die Wahrscheinlichkeitsmasse auf wenige Komponenten verteilt ist. Das führt zu folgenden Aussagen über die Modellbeschaffenheit und Eigenschaften der Topic-Modelle:

- Über den Parameter der Dirichlet-Verteilungen wird gesteuert, ob die Komponenten im Modell relativ gleich behandelt werden oder ob ein Fokus auf wenige Themen gelegt werden soll. Im Fall der Verteilung $p(\mathbf{w}|z)$, wirkt sich der Parameter auf die Spezifizität eines Themas aus, da die Verteilung der Wahrscheinlichkeitsmasse auf die einzelne Terme in einem Thema beeinflusst wird.

- Der Parameter β repräsentiert in der LDA die a priori Annahme über die Verteilung der Wörter innerhalb der Themen im Modell. Werden Verteilungen $p(\mathbf{w}|z)$ mit sehr hoher Wahrscheinlichkeitsmasse auf wenigen Types erwartet, so muss das Modell mit mehr Themen K inferiert werden, da mehr Themen benötigt werden, um die einzelnen Wahrscheinlichkeiten der Wörter optimal zu verteilen.

Werden jedoch gleichmäßige Wahrscheinlichkeiten für alle Wörter in einem Thema erwartet, wie bei einem Parameter $\alpha > 1$, so müssen die Wörter auf weniger Themen verteilt werden, um das Modell mit einer hohen Wahrscheinlichkeit zu berechnen. Die Themen werden allgemeiner und abstrakter.

Um dieses Verhalten zu demonstrieren, kann die Veränderung der Modelleigenschaften bei einer Veränderung des β Parameters in einem HDP-LDA Modell erprobt werden. Im Fall der HDP-LDA, wird die Verteilung $p(\mathbf{w}|z)$ durch den Parameter β beeinflusst und somit auch der Abstraktionsgrad der Themen. Je geringer der Wert gewählt wird, desto spezifischer sind die Themen und ein großer Anteil der Wahrscheinlichkeitsmasse in $p(\mathbf{w}|z)$ wird auf wenige Terme verteilt. Für die Erprobung der Hyperparameters β wird ein HDP-LDA Modell gewählt, da es die optimale Anzahl der Themen automatisch findet und den Einfluss der Verteilung $p(\mathbf{w}|z)$ auf die Themenanzahl gut reflektiert. Im Experiment wird die resultierende Themenanzahl in einem Modell für unterschiedliche Parametereinstellungen von β ermittelt, während alle anderen Parameter stabil gehalten werden. In Abbildung 3.7 ist dargestellt, wie die Anzahl der Themen abnimmt, je allgemeiner und gleichverteilter $p(\mathbf{w}|z)$, in Abhängigkeit vom Parameter β, wird. Das Verhalten ist nicht linear und K wird in einem Bereich von 0.01 - 0.2 sehr stark beeinflusst. In einem Wertebereich $\beta > 0.2$ bleiben die resultierenden K nahezu konstant. In dem hier dargestellten Experiment wird ein HDP-LDA Modell mit einem Chinese-Restaurant-Franchise Gibbs Sampler inferiert (Teh u. Jordan, 2010). Da für die Inferenz ein Gibbs-Sampler verwendet wird, der jedem Wort ein Thema zuordnet, kann gemessen werden, wie viele Token mit einem Thema assoziiert sind. In Abbildung 3.8 ist zu erkennen, dass bei weniger Themen mehr Wörter in den jeweiligen Themen zugeordnet werden müssen. Die Parameter der Dirichlet Verteilung sind im LDA und HDP-LDA Modell demnach essentiell für die Festlegung der Themengranularität und der damit einhergehenden Themeninterpretation. Diese Festlegung entscheidet gleichzeitig, wie viele Themen in einem Text zu finden sind. Es kommt darauf an, in welchem Detail die Themen einer Textkollektion extrahiert werden sollen.

Unter den Verfahren, die unter die Topic-Modelle fallen, existieren allerdings auch andere Ansätze, die auf anderen Verteilungsannahmen basieren. So kann beispielsweise eine Poisson Verteilung angenommen werden (Teh, 2006; Teh u. Jordan, 2010), welche die Dirichlet ersetzt. Für diese Erweiterungen und Modellannahmen müssen die Parameter der zugrundeliegenden Verteilungen in ihrer Bedeutung für das Modell verstanden werden und entsprechend eingestellt werden. In den weiteren Ausfüh-

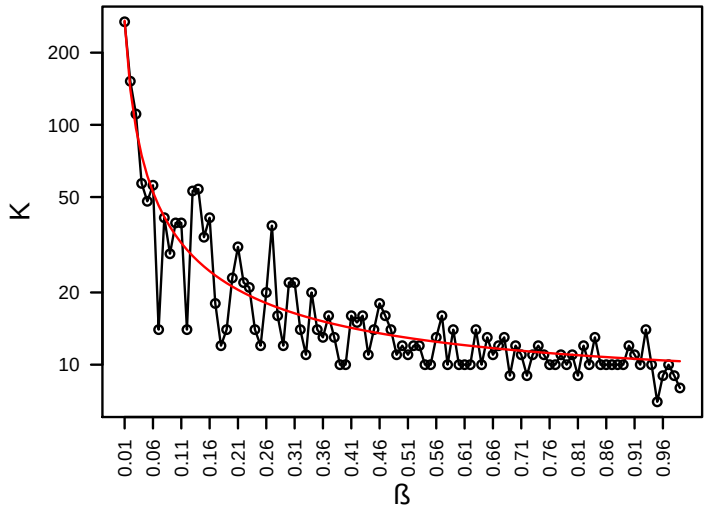

Abbildung 3.7: Anzahl der Themen K in Abhängigkeit vom Parameter β, in einem HDP-LDA Modell. Die Achse für K ist log-skaliert.

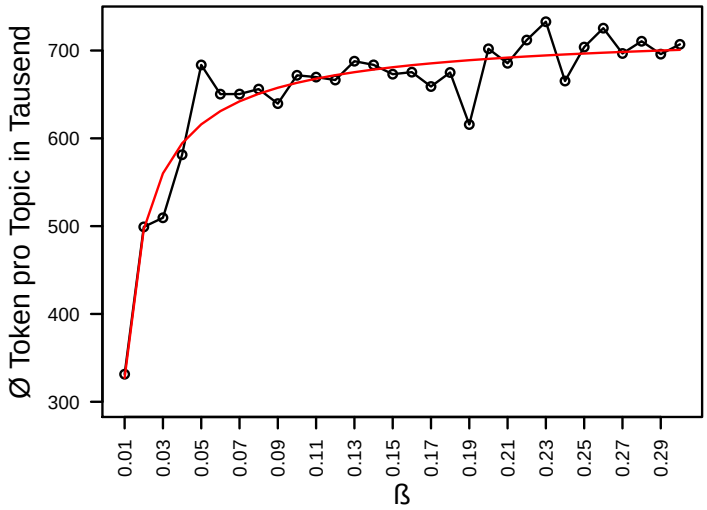

Abbildung 3.8: Darstellung der durchschnittlichen Anzahl von Token, die einem Thema in einem HDP-LDA Modell, in Abhängigkeit des Parameters β, zugeordnet werden.

rungen konzentriert sich die Arbeit dennoch auf die Modelle, die mit der Dirichlet Verteilung arbeiten, um die Nachvollziehbarkeit der Untersuchungen zu gewährleisten.

Die Einstellung einzelner Parameter ist, wie am Beispiel der Latent Dirichlet Allocation, immer ein Zusammenspiel aller verwendeten Parameter. Soll beispielsweise bei der Veränderung der Verteilungen $p(\mathbf{w}|z)$, durch den Hyperparameter β, die Wahrscheinlichkeit des Modells innerhalb der Daten maximiert werden, muss der Modellparameter K angepasst werden. Dies passiert beim HDP-LDA Modell automatisch. Das Zusammenspiel ist im Fall der LDA schwer einzuschätzen und schwer zu evaluieren, da die Berechnungszeit der Gibbs Sampler für jedes Modell in großen Korpora sehr hoch sein kann. Es ist deshalb nicht praktikabel, dass verschiedene Parametereinstellungen evaluiert werden. Es ist praktischer, wenn nur die zu extrahierende Anzahl der Themen festgelegt werden muss und die Parameter für α und β auf ein Optimum eingeschätzt werden können. Für viele Topic-Modelle und deren Inferenzverfahren wurden solche Schätzer bereits entwickelt (Blei u. a., 2003; Heinrich, 2005; Hoffman u. a., 2013). Mit Inferenzen, die in der Lage sind, Parameter zu optimieren, besteht ferner die Möglichkeit, einen Parameter zu fixieren und andere zu optimieren. Meist werden die Hyperparameter in Gibbs Samplern nach mehreren Iterationen neu berechnet, sodass die Modellwahrscheinlichkeit maximiert wird. So kann die Modellwahl an das Analyseproblem angepasst werden, indem tatsächlich nur die Parameter fest vergeben werden, die aktiv das gewünschte Verhalten einer Analyse beeinflussen. Bei der LDA ist dies in den meisten Fällen die Anzahl der zu extrahierenden Themen K. Bei einer automatischen Bestimmung der Themenanzahl, wie im HDP-LDA Modell, kann die Dichte der Verteilung $p(\mathbf{w}|z)$ durch einen Hyperparameter bestimmt werden, sodass die Abstraktion der Themen bestimmt wird, nicht die Anzahl der Themen. Bei der externen Festlegung von K kann es passieren, dass durch die Schätzung der Hyperparameter die Auflösung der resultierenden Themen nicht den Erwartungen entspricht. Das Modell muss mit einer anderen Themenanzahl noch einmal gerechnet werden. Dies ist aber immer schneller und praktikabler, als alle möglichen Kombinationen von Themen und Hyperparametern zu erproben, ohne diese automatisch zu schätzen. Somit soll die Analyse immer mit der Festlegung eines Hyperparameters durchgeführt werden, der das Ziel der Analyse, beispielsweise Abstraktion oder Themenanzahl, reflektiert. Anhand dieses Parameters wird entschieden, ob die Analyse erfolgreich ist oder mit einem anderen Parameterwert wiederholt werden muss. Mit dieser Arbeitsweise müssen weniger Probemodelle und Tests durchgeführt werden. Weiterhin können die Auswirkungen eines einzelnen variablen Hyperparameters besser verstanden werden, wenn die anderen Parameter immer im Zusammenspiel

optimal eingeschätzt werden. Es empfiehlt sich deshalb immer, automatische Parameterschätzer in den Inferenzen einzusetzen. Nur so kann das Ergebnis und dessen Beeinflussung durch einzelne Parameter korrekt eingeschätzt werden. Im Gegensatz zum Schwellwertparameter im TDT, können keine optimalen Werte für die Parameter der Topic-Modelle empirisch getestet und vorgeschlagen werden. Für den Einsatz von Topic-Modellen in Inhaltsanalysen empfiehlt sich deshalb das folgende methodische Vorgehen, um die Konfiguration der Parameter vorzunehmen.

1. Ankodierung (Modellberechnung) einer kleineren Dokumentmenge oder weniger Zeitscheiben, falls die Gesamtmenge zu groß ist, mit einer initialen Parametereinstellung.
2. Abgleich der Themenauflösung und der möglichen Interpretation mit den Analyseanforderungen, die aus Vorannahmen oder der Definition der Inhaltsanalyse erstellt werden.
3. Iterative Anpassung der Parameter bis zur gewünschten Ergebnisdarstellung und validen Interpretation hinsichtlich der Analyseanforderung.
4. Vollständige Modellberechnung auf den Daten.
5. Durchführung der Interpretation und Auswertung der Ergebnisse.

3.2.4 Anwendung

Das Verfahren stellt ebenfalls einen Bag-of-words Ansatz dar, sodass die Dokumente als Vektor über das Vokabular dargestellt werden müssen, wie es auf Seite 45 gezeigt wird. Ähnlich wie bei der Clusterung im TDT-Verfahren, kann auch bei Topic-Modellen ein algorithmisches Vorgehen für die Verwendung vorgeschlagen werden. Die Modelle sind allerdings nicht in der Lage, neue Dokumente in die Modellberechnung mit einzubeziehen. Das heißt, dass die Modelle immer mit einer abgeschlossenen Dokumentmenge arbeiten und deshalb nicht für permanent eingehende Datenströme oder sequenzielle Batches eingesetzt werden können. Für die Berechnung mehrerer Tage oder Jahre ist es nicht möglich, die Dokumentmenge aufzuteilen und sequenziell in einem Modell zu bearbeiten. Dies hat den Nachteil, dass die Berechnung großer Dokumentmengen immer komplett erfolgen muss. In großen Dokumentmengen, wie beispielsweise ganzen Jahrgängen von Nachrichtenartikeln, sind sehr viele Themen enthalten, deren vollständige Auswertung für Inhaltsanalysen sehr komplex wäre. Für eine rein technische Lösung, wie die Realisierung einer Dokumentsuche, ist dies nicht

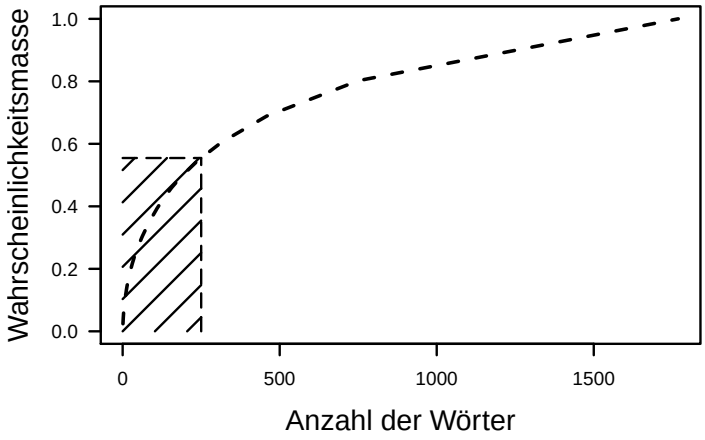

Abbildung 3.9: Kumulierte Dichte der Wahrscheinlichkeiten der Terme in einem Thema. Ein überwiegender Teil der Masse liegt bei wenigen Termen.

von Belang, da die Themen hier nicht interpretiert werden müssen. Es ist für Inhaltsanalysen deshalb einfacher, die Zerlegung in die Themen anhand kleiner Datenmengen vorzunehmen. Natürlich sprechen diese Aussagen nicht dagegen, ein Topic-Modell für eine Analyse großer Datenmengen einzusetzen. Für die Bearbeitung diachroner Quellen, die oft mehrere zehntausende Dokumente enthalten, wäre es dennoch hilfreich, Topic-Modelle sequenziell zu berechnen, um für neue Daten und die getrennte Analyse einzelner Zeitscheiben offen zu bleiben. Modelle für einzelne Zeitscheiben müssen in Zusammenhang gebracht werden, sodass verschiedene Zeitscheiben vergleichbar sind.

Für die Einzelberechnung von Topic-Modellen in Zeitscheiben und die Vergleichbarkeit der Themenstrukturen in den einzelnen Zeitscheiben ist es wichtig, dass gleiche oder ähnliche Themen verschiedener Modelle in verschiedenen Zeitscheiben assoziiert werden können. Daraufhin können übereinstimmende Themen untereinander verbunden und zugeordnet werden. Für den Vergleich unterschiedlicher Modelle wird in Niekler u. Jähnichen (2012) vorgeschlagen, die Verteilungen $p(\mathbf{w}|z)$ der Modelle direkt zu vergleichen. Für diesen Zweck werden verschiedene Distanzmaße für den Vergleich der Verteilungen erprobt. Im Folgenden wird diese Idee für die Anwendung in Themenanalysen genauer erläutert. In unterschiedlichen Zeitscheiben oder Quellen werden Themen mit anderen Aspekten dargestellt. Deshalb kann davon ausgegangen werden, dass die berechneten Themen aus verschiedenen Topic-Modellen, obwohl Sie thematisch identisch sind, in einem Vergleich der Verteilungen nicht vollständig übereinstimmen. Der vorgeschlagene Vergleich nutzt die Eigenschaft der Verteilung $p(\mathbf{w}|z)$, dass viel Wahrscheinlichkeitsmasse auf wenigen Termen liegt. Der Zusammenhang ist

in Abbildung 3.9 dargestellt. Diese wenigen Terme repräsentieren einen statistischen Verwendungszusammenhang, der sich als Sinn oder Inhalt des Themas interpretieren lässt. In Topic-Modellen wird allen Wortarten in einem Korpus in jedem Thema eine Wahrscheinlichkeit > 0 zugewiesen. Dadurch ist es schwer, anhand aller Wörter zu entscheiden, ob ein Thema in einem anderen Topic-Modell inhaltlich ähnlich ist. Diese Schwierigkeiten können durch eine Veränderung der Verteilung umgangen werden. Durch Abschneiden der Terme mit einer geringen Wahrscheinlichkeit wird erreicht, dass nur die Essenz der Themen miteinander verglichen wird. Die Experimente zu dieser Idee werden so definiert, dass zunächst alle Wörter in den Verteilungen nach ihrer Themenwahrscheinlichkeit absteigend sortiert werden. Anschließend werden die n Terme mit der höchsten Wahrscheinlichkeit belassen und alle anderen Wahrscheinlichkeiten in einem Thema auf den Wert 0 gesetzt. Die verbleibenden Werte können als Vektor betrachtet werden, der aus der Multinomialverteilung erzeugt wird. Diese Vektoren können mit unterschiedlichen Ähnlichkeitsmaßen verglichen werden. Es wird für verschiedene n und verschiedenen Ähnlichkeitsmaße erprobt, welche Kombination am effektivsten arbeitet. Es soll möglichst große Ähnlichkeit für gleiche Themen und möglichst kleine Ähnlichkeit für unterschiedliche Themen erzeugt werden. Für das Experiment werden ähnliche Themen aus unterschiedlichen Zeitscheiben der Online-Ausgabe des Guardian verwendet. Die Themen für die Experimente sind in Tabelle 3.5 dargestellt. Um festzustellen, welches Ähnlichkeitsmaß geeignet ist, wird festgelegt, dass gleiche Themen theoretisch eine Ähnlichkeit von 1 aufweisen müssen, wenn davon ausgegangen wird, dass die Ähnlichkeit als Maß zwischen 0 und 1 ausgedrückt wird. Analog sollen unterschiedliche Themen unter diesen Bedingungen eine Ähnlichkeit von 0 aufweisen. Anschließend werden geeignete Maße wie Jenson-Shannon (JS), Dice und das Kosinusmaß erprobt (Niekler u. Jähnichen, 2012). Die Maße Dice und JS werden zu Ähnlichkeitsmaßen, wenn die resultieren Werte von 1 subtrahiert werden. Die vorgeschlagenen Maße sind alle auf einen Wertebereich zwischen 0 und 1 beschränkt. In den Experimenten wird bestimmt, welche Ähnlichkeit tatsächlich mit einem der Maße und einer festgelegten Anzahl n der Terme gemessen werden kann. Für jeden Vergleich der vorher ausgewählten Themen aus Tabelle 3.5 wird bestimmt, wie der errechnete Wert vom erwarteten Wert abweicht. Wie gut ein Maß unter der Auswahl von n hochwahrscheinlichen Termen die Ähnlichkeit abbildet, wird durch die mittlere Abweichung (Mean Deviation) aller gemessenen Paarungen

$$MD = \frac{1}{N_{topics}^2} \sum_{i=1}^{N_{topics}} \sum_{j=1}^{N_{topics}} \|s_{ij} - s_{ij}^*\|, \tag{3.7}$$

Datum	Kurzname	Top 20 Worte
12-03-2011	japan1	Japan nuclear plant tsunami earthquake reactor power Japanese disaster radiation water damage quake plants country Tokyo explosion reactors Fukushima reports
13-03-2011	japan2	nuclear Japan tsunami power earthquake reactor Japanese water disaster plant radiation crisis plants magnitude fuel reactors aftershocks rescue Friday prefecture explosion
14-03-2011	japan3	nuclear Japan reactor power plant Japanese earthquake tsunami explosion disaster Tokyo rescue reactors energy plants crisis radiation JST safety water
15-03-2011	japan4	nuclear Japan plant power radiation Japanese reactor reactors fuel earthquake levels Tokyo water disaster tsunami fire level crisis agency safety
10-03-2011	libya1	Libya Gaddafi forces military zone no-fly Nato Libyan Libyan oil foreign rebels rebel council Ras_Lanuf France fighting regime defence country
12-03-2011	libya2	Gaddafi Benghazi MP country regime revolution revolutionary Libya forces GG international council countries intervention foreign eurozone Libyan no-fly city army
13-03-2011	libya3	Gaddafi Libya oil foreign Arab Europe intervention no-fly Iraq zone support military forces regime rebels security western uprising Egypt Tunisia
14-03-2011	libya4	Cameron Labour Libya zone Gaddafi no-fly Miliband Balls Britain vote tax campaign action plan party Clegg ministers Labour rebels referendum
15-03-2011	libya5	no-fly zone Bahrain forces Gaddafi military Libya troops security rebels foreign torture regime Benghazi told Saudi_Arabia Britain France G8 town
15-03-2011	stopwords1	years public make work pay world made good UK back part long ve don day Germany week big report

Tabelle 3.5: Ausgewählte Themen, erzeut durch eine LDA, der Online-Ausgabe des Guardian im Zeitraum vom 10. - 15. März 2011.

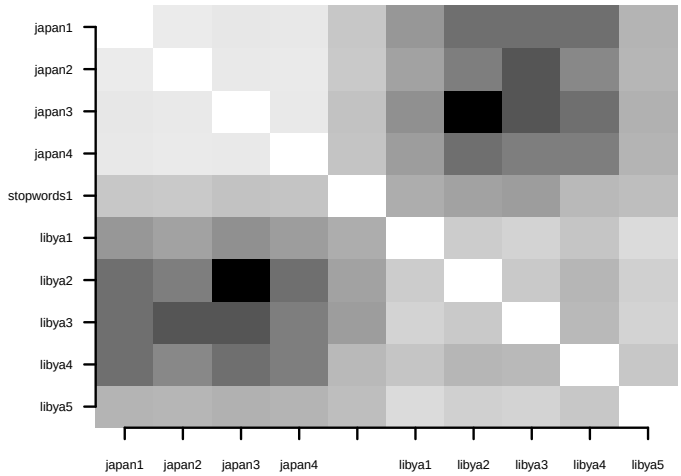

Abbildung 3.10: Darstellung der Ähnlichkeitsmatrix mit dem Maß JS. Die Helligkeit der Zellen gibt die Ähnlichkeit der Themen an. Je dunkler die Zellen erscheinen, desto kleiner ist die Ähnlichkeit.

mit N_{topics} der Anzahl der Themen, die verglichen werden, s_{ij} der gemessenen Ähnlichkeit und s^*_{ij} die erwartete Ähnlichkeit, ermittelt. Je geringer die mittlere Abweichung ist, desto besser ist ein Maß und die Unterauswahl der Terme geeignet. Die Experimente zeigen, dass unter Verwendung des Kosinusmaß ca. 10-40 Terme ausreichen, um die Distanz optimal an die Erwartung anzupassen (Niekler u. Jähnichen, 2012). In Abbildung 3.12 werden die Ergebnisse des Experiments dargestellt. Um noch einmal zu verdeutlichen, wie sich die Ähnlichkeiten innerhalb der Beispielthemen aus Tabelle 3.5 verhalten, werden in den Abbildungen 3.10 und 3.11 die Ähnlichkeiten, durch Einfärbung der Zellen in der Ähnlichkeitsmatrix, visualisiert. In der auf dem Maß JS basierenden Darstellung ist zu erkennen, dass die Verwendung aller Therme in den Themen zu kaum unterscheidbaren Ähnlichkeiten führt. Werden jedoch die Ergebnisse des Experiments eingesetzt und nur 40 Therme für den Vergleich mit dem Kosinusmaß verwendet, so ergeben sich hohe Ähnlichkeiten für inhaltlich verwandte Themen und eine klare Abgrenzung unter nicht zusammengehörigen Themen.

Die Übereinstimmung zweier Themen in unterschiedlichen Quellen kann durch die Kosinusdistanz

$$s(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{3.8}$$

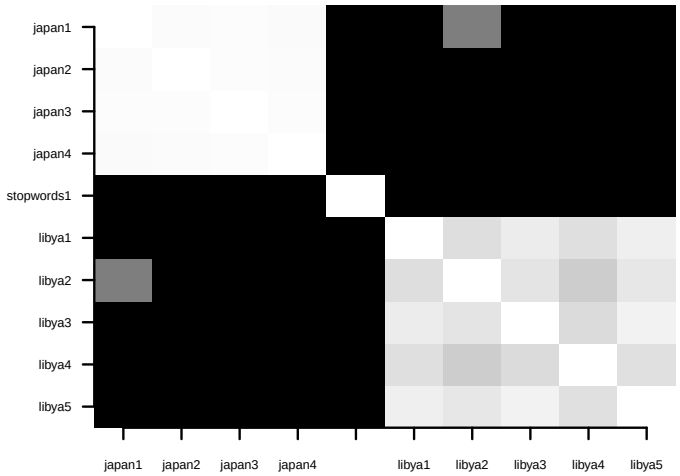

Abbildung 3.11: Darstellung der Ähnlichkeitsmatrix mit dem Kosinusmaß und einer Reduktion der verwendeten Terme auf 40. Die Helligkeit der Zellen gibt die Ähnlichkeit der Themen an. Je dunkler die Zellen erscheinen, desto kleiner ist die Ähnlichkeit.

bestimmt werden. Die Größen $p(\mathbf{w}|z_k)$ und $p(\mathbf{w}|z_k^*)$ werden als Vergleichswerte A und B gesetzt. Durch das Setzen eines Schwellwertes für die Ähnlichkeit, der anhand von Probemessungen ermittelt wird, können Themen unterschiedlicher Dokumentkollektionen oder Zeiträume verglichen und als identisch betrachtet werden. Dadurch kann ein Tracking der Themen mit Hilfe von Topic-Modellen in unterschiedlichen Zeitscheiben betrieben werden. In aufeinanderfolgenden Zeitscheiben werden jeweils Topic-Modelle berechnet und die Themen mit der vorgeschlagenen Vergleichsprozedur verbunden. Ähnlich wie bei TDT kann damit ein Algorithmus gefunden werden, der die Berechnung und Zuordnung der Themen zueinander in den Zeitscheiben durchführt. Ein Besonderheit ist, dass in einer Zeitscheibe bzw. einem Batch ein Thema aus einer vergangenen oder älteren Zeitscheibe nur einmal zugeordnet werden kann. Ansonsten würde ein solches Vorgehen erlauben, dass sich ein Thema aufspalten kann. Dies kann sinnvoll sein, wenn genau dieses Verhalten beobachtet werden soll. In der hier gezeigten Anwendung soll die Aufspaltung und Vereinigung von Themen vermieden werden. Aus diesem Grund ist in Prozedur 2, die im Anhang zu finden ist, in Zeile 14 bzw. Zeile 22 eine Bedingung implementiert, die ein doppeltes Zuordnen von Themen aus vorangegangenen Zeitscheiben verhindert. Auf diese Art und Weise können sowohl statisch diachrone Textkollektionen als auch erweiterbar diachrone Textkollektionen geclustert werden. Die Anzahl der Vergleiche wächst linear mit den Zeitscheiben, die in die Vergleiche einbezogen werden. Es ist, wie bei TDT sinnvoll, die Anzahl der einbezogenen Zeitscheiben zu begrenzen. Die Auswirkungen auf das

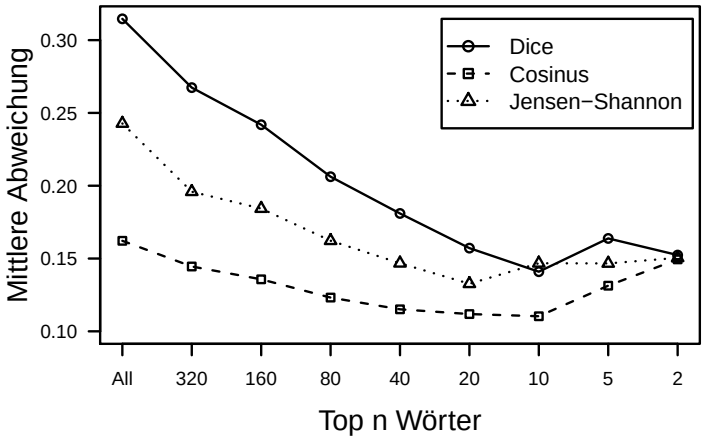

Abbildung 3.12: Abtragung der mittleren Abweichung von der erwarteten Distanz für den Themenvergleich mit unterschiedlichen Maßen. Die Abweichung ist für verschiedene Distanzmaße und verschiedene n, die Einbeziehung der wahrscheinlichsten Terme, abgetragen.

Laufzeitverhalten mit unterschiedlichen Horizonten sind in Abbildung 3.13 dargestellt. Das beschriebene Verfahren ist ähnlich zur Arbeit von Kim u. Oh (2011). Allerdings wird dort nicht die Anzahl der Worte reduziert, die für den Vergleich herangezogen werden. Aus diesem Grund schneidet JS in Kim u. Oh (2011) besser ab, da es für Wahrscheinlichkeitsverteilungen das geeignetere Maß ist. Wird jedoch das Vokabular für den Vergleich reduziert, so ergibt sich ein anderes Bild. Die hier gezeigte Vorgehensweise und die Strategie unterscheiden sich, trotz Ähnlichkeiten, zu den Beiträgen von Kim u. Oh (2011). Das hier vorgestellte Verfahren soll möglichst semantisch ähnliche Themen finden. Wenn sich ein Thema inhaltlich sehr stark verändert, so soll ein neues Thema erfasst werden. Der hier vorgeschlagene Mechanismus soll die Distanzen semantisch ähnlicher Themen künstlich verringern, bzw. semantisch nicht ähnlicher Themen erhöhen, sodass eine klare Trennung vorliegt. Werden alle Terme eines Themas als Grundlage für die Vergleiche herangezogen, so besteht durch die Eigenschaften der Multinomialverteilungen der Topic-Modelle grundsätzlich immer eine gewisse Ähnlichkeit.

Wie bereits erwähnt, kann ein Topic-Modell auf der kompletten Dokumentmenge berechnet werden. Dies stellt gewissermaßen eine „globale“ Modellberechnung dar. Zusammenfassend können Topic-Modelle auf zwei verschiedene Arten genutzt werden:

- Durch die Brechung einer diachronen Textkollektion im ganzen und

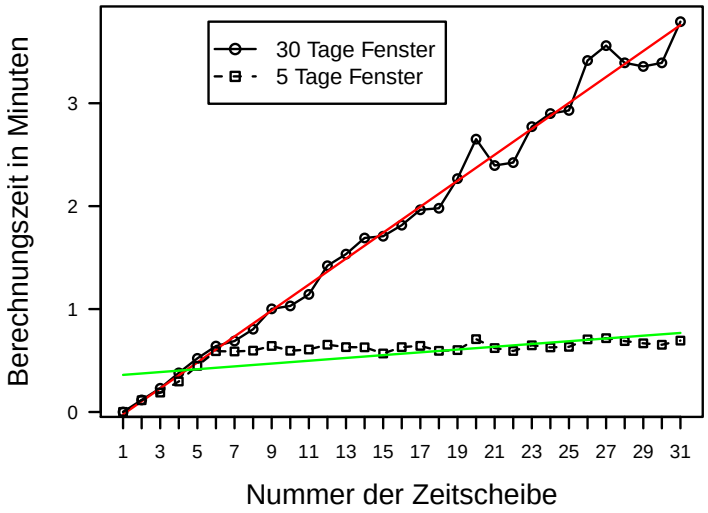

Abbildung 3.13: Laufzeitverhalten zur Berechnung der Batches mit Topic-Modellen uns Algorithmus2 für unterschiedliche retrospektive Horizonte über Tageszeitscheiben.

- durch die Zerlegung einer Kollektion in unterschiedliche Batches, die aus Zeitscheiben erstellt werden können.

Wenn eine abgeschlossene Kollektion vorliegt, so kann die erste Variante durchaus genutzt werden. Aber das LDA oder HDP-LDA Modell berücksichtigt die Zeitabhängigkeiten nicht. Eine Ausnahme bilden Dynamic-Topic-Modelle, die allerdings auch mit abgeschlossenen Kollektionen arbeiten und deshalb nicht für das Tracking, sondern für retrospektive Analysen geeignet sind. Ist die Kollektion nicht abschlossen oder soll ein Tracking der Themen erfolgen, so muss die zweite Variante gewählt werden, da Topic-Modelle abgeschlossene Datenmengen zur Berechnung erfordern. Der Rechenaufwand wird immer höher je mehr eine Textkollektion wächst oder je größer sie ist. Soll die Entwicklung oder der Trend der Themen innerhalb einer diachronen Kollektion verfolgt werden, so muss mit Zeitscheiben gearbeitet werden. Bei globalen Modellen können die Themenanteile anhand der Zeitstempel der Dokumente verortet werden. Die Berechnungen sind allerdings nicht offen für neue Daten und die Themenmengen können sehr groß und unübersichtlich sein. Bei der Konzentration auf kleinere Einheiten, wie Tage oder Monate, ist eine qualitative Kontrolle und Einordnung der Ergebnisse unproblematischer.

In beiden Fällen ist das Ergebnis einer Topic-Modell Analyse eine Menge von Themen K. Den Dokumenten sind die Themen mit bestimmten Anteilen $p(\mathbf{z}|d)$ zugeordnet. Jedes Dokument kann als eine Verteilung der Themen innerhalb des Dokuments beschrieben werden. Die Themen werden als Verteilungen über das Vokabular, $p(\mathbf{z}|z)$,

dargestellt. Für eine thematische Zusammenfassung der Inhalte eines Korpus oder einer Zeitscheibe können die Themenverteilungen herangezogen werden. Die Terme können aufsteigend nach ihrer Wahrscheinlichkeit innerhalb einer Themenverteilung sortiert werden. Als Zusammenfassung wird eine Auswahl der wahrscheinlichsten Terme bestimmt, da diese den hauptsächlichen semantischen Zusammenhang eines Themas in einem Topic-Modell repräsentieren. Beispielhaft wird dies in Tabelle 3.4 und 3.5 gezeigt. Anders als bei TDT kann aus den Verteilungen in den einzelnen Dokumenten kein Durchschnitt errechnet werden, um Dokumente oder Cluster zusammenzufassen. Dennoch können die Verteilungen, insbesondere die Verteilungen $p(\mathbf{z}|d)$ und $p(\mathbf{w}|z)$, genutzt werden, um Abstraktionen und Zusammenfassungen für eine Dokumentmenge oder Einzeldokumente darzustellen. So kann die Verteilung $p(\mathbf{z}|d)$ genutzt werden, um die wichtigsten Themen eines Dokuments zu finden. Die Menge aller Dokumente D_k, die einem Thema zugeordnet sind, kann bestimmt werden, indem ein Schwellwert festgelegt wird, ab welchem Themenanteil $p(\mathbf{z}|d)$ ein Dokument einem Thema k zugeordnet werden kann.

Die linguistischen Theorien über die Repräsentation von Themen sind demnach ähnlich anschlussfähig, wie beim TDT-Verfahren. Die Themenzusammenfassungen aus den Verteilungen können als Nominalgruppen interpretiert werden. Natürlich enthalten die Themenverteilungen auch andere Wortarten. In der Betrachtung der wichtigsten Terme eines Themas, tragen jedoch hier die Nomen, Ereignisbezüge und Eingennamen zentral zur Bedeutung eines Themas bei. Abweichend von der Forderung, dass eine Nominalgruppe unsortiert dargestellt wird, kann durch die klare Zuweisung von Wahrscheinlichkeiten eine Sortierung vorgenommen werden. Dies erlaubt eine Interpretation darüber, welche Terme wichtiger und präsenter sind. Dennoch bleibt es bei der Dokument- und Kollektionsdarstellung schwierig, die Zusammenhänge zwischen den einzelnen Begriffen innerhalb einer Themenbeschreibung zu interpretieren, da die propositionalen Inhalte eine Themas nicht erfasst werden, sondern lediglich distributionelle Eingenschaften.

Eine entscheidender Vorteil beim Einsatz von Topic-Modellen ist die Möglichkeit, dass Dokumente mehreren Themen zugeordnet werden können. Somit können Nebenthemen und thematische Sprünge innerhalb von Dokumenten erkannt werden. Die Themen sind, im Gegensatz zu den Zusammenfassungen der Cluster aus dem TDT-Verfahren, eine Auswahlfunktion, derer sich die Dokumente bedienen. Dies zeigt eine Anschlussfähigkeit an die Fokustheorie, die Themen als Auswahlfunktion beschreibt. Eine Beschränkung besteht aber darin, dass die Absichten der Themen oder Texte nicht durch Topic-Modelle erfasst werden können. Ein Ziel hinter einem Text ist

durch die Betrachtung der darin enthaltenen Themen höchstens zu erraten oder zu interpretieren, da die Wortverteilungen keinerlei Möglichkeit bieten, diese psychologische Komponente abzubilden. Dennoch ist eine diachrone Textkollektion nicht nur als Sammlung unterschiedlicher Dokumente ähnlicher inhaltlicher Struktur oder Cluster abbildbar, sondern kann durch mehrere inhaltliche Faktoren aus denen die Dokumente zusammengesetzt sind, beschrieben werden. Einzelne Dokumente werden nicht als mono-thematisch vorausgesetzt. Dennoch bietet sich bei der Verwendung von Topic-Modellen nicht die Möglichkeit, dass Konzepte wie die thematische Progression (Thema-Rhema) oder die Bildung von Makropropositionen abgebildet werden können. Da diese Konzepte sehr stark auf die Analyse einzelner Dokumente bezogen sind, widerspricht dies im Grunde den statistischen Herangehensweisen der Topic-Modelle und deren globaler Sicht auf die Dokumente.

3.3 Signifikante Kookkurrenzen

Die bisher vorgestellten Ansätze zur Themenanalyse eignen sich unter bestimmten Voraussetzungen und Einschränkungen zu einer Themenanalyse in Sinne linguistischer und inhaltsanalytischer Sichtweisen. Insbesondere kann die Bildung thematischer Kategorien und Dokumentmengen durch Clusterung und distributionelle Eigenschaften der Texte empirisch unterstützt werden. Auf diese Art und Weise können Themenstrukturen in großen Textkollektionen extrahiert werden. Gewissermaßen bilden die bisherigen Verfahren die Textinhalte ab, die einer geringen Dynamik unterliegen und die thematische Grundlage für die Textproduktion repräsentieren.[9] In diesen Themenanalysen ist es allerdings schwer möglich, Propositionen unterschiedlicher Abstraktion aus den Themendarstellungen abzulesen, da die Themendarstellungen keine propositionale Struktur, Wortreihenfolge oder Satzbezüge aufweisen. Die neuen Aussagen (hohe Dynamik) über ein Thema und die Entfaltung eines Themas bleiben bei der Anwendung der Verfahren nicht direkt greifbar. Die beteiligten Wörter finden sich zwar in den Textzusammenfassungen, deren Zusammenhang und Kontext ist durch die bloße distributionelle Verteilung der Terme und deren Darstellung als Wortliste schwer zu interpretieren. In der Themenanalyse nach van Dijk (1980) und Mackeldey (1987) ist dies aber eine grundsätzliche Forderung, um Themen beschreiben und verstehen zu können. Auch Arbeiten zur empirischen Kategorienbildung in der Kommunikationswissenschaft beziehen sich auf die propositionale Struktur der Themen (Früh (2001), Früh (2007, vgl. S. 273 ff.)). Die Abstraktion und Ableitung allgemeiner

[9] Auf die Dynamik von Wörtern wird in Abschnitt 2.1.3 auf Seite 29 eingegangen.

semantischer Regeln, die Bildung von Oberbegriffen oder die Zusammenfassung von Termen ähnlicher Bedeutung ist mit Topic-Modellen und TF/IDF basiertem Clustering nicht möglich. Denn soll zusätzlich zur Analyse des Themengehalts auch eine genaue Zusammenfassung der Themenbedeutung angefertigt werden, so sind mehr Informationen über die Bedeutung der Wörter, die ein Thema enthält, nötig. Um diese statistisch zu erfassen, eignet sich eine Kookkurrenzanalyse der Dokumente, die einer Thematisierung zugeordnet werden können. Diese Anwendung ist von einer globalen Kookkurrenzanalyse zu unterscheiden, die alle Dokument einer Kollektion einbezieht. Bei einer globalen Anwendung vermischen sich die Bedeutungen mehrerer Thematisierungen, wenn das Wort eine thematisch ambige Verwendung hat. Deshalb ist in der hier definierten Form der Kookkurrenzanalyse immer eine Dokumentmenge D_k als Ausgangspunkt definiert, die sich auf eine Thematisierung k bezieht. Dadurch bildet eine Kookkurrenzanalyse nur einen thematischen Sinnzusammenhang ab. Wie eine solche Dokumentmenge aus den bereits vorgestellten Verfahren zur Themenanalyse gebildet werden kann und definiert wird, ist in Abschnitt 3.4 genauer dargestellt.

Bisher fehlt der vorgestellten Methodik die Möglichkeit, Propositionen aus statistischen Gemeinsamkeiten mehrerer Dokumente oder Textstellen, die einem Thema zugeordnet sind, abzuleiten. Die Betrachtung syntagmatischer und paradigmatischer Relationen innerhalb der Kontexteinheiten eines Themas, muss deshalb herangezogen werden, um die semantischen Bedeutungen der Terme erklären zu können. Eine syntagmatische Beziehung zwischen zwei Worten besteht, wenn diese häufig gemeinsam in einer Kontexteinheit auftreten (Saussure, 2001; Heyer, 2006; Busse, 2009). Als paradigmatische Relation gelten diejenigen, bei denen Terme in gleichen Kontexten verwendet werden (Saussure, 2001; Heyer, 2006). Die Semantik, als Bedeutung der Wörter, betrachtet Einzelwörter durch ihre gemeinsame Verwendung mit anderen Wörtern. Um die semantischen Relationen innerhalb einer Menge von Texten zu erfassen, müssen die paradigmatischen und syntagmatischen Relationen analysiert werden. Dabei stehen die statistisch-syntagmatischen Relationen im Vordergrund. Diese können über eine Analyse signifikanter Kookkurrenzen aus dem Text destilliert werden. Die Kookkurrenz zweier gemeinsam auftretender Wortformen in einer Kontexteinheit wird auf der Grundlage eines Signifikanzmaßes berechnet(Heyer, 2006, vgl. S. 24). Demnach stehen zwei Wortformen nur in syntagmatischer Relation, wenn ihr gemeinsames Auftreten nicht zufällig ist. Das statistisch signifikante gemeinsame Auftreten wird als signifikante Kookkurrenz bezeichnet und immer innerhalb eines lokalen Kontexts gemessen (Heyer, 2006, vgl. S. 23). Der lokale Kontext ist die Satz-, Absatz- oder Dokumentebene (Manning, 1999, vgl. S. 297) und entspricht in der

inhaltsanalytischen Sichtweise der Kontexteinheit. Die Signifikanz wird nach einem zu bestimmenden Maß errechnet und vergleicht das gemeinsame Auftreten mit einer erwarteten Häufigkeit zufällig gewählter Wortformen (Heyer, 2006, vgl. S. 136). Als statistische Grundlage dienen alle Kontexteinheiten, die für die Zählung der Kookkurrenzen einbezogen werden und den globalen Kontext in Bezug auf die zur Verfügung stehende Textmenge abbilden. Aus den signifikanten Kookkurrenzen lassen sich Wortverwendungen als „wortübergreifende sprachliche Einheiten“ (Busse, 2009, vgl. S. 112) der lokalen Kontexte ablesen. Dies lässt jedoch offen, in welcher Weise die Begriffe semantisch verbunden sind und eine ähnliche Interpretation, wie bei den bisher vorgestellten Themenbetrachtungen ist die Folge, da die Relationen zwischen den Wortformen keine Richtungen oder logische Verknüpfungen aufweisen. Grundsätzlich lassen sich statistische Zusammenhänge und die Entwicklung dieser Zusammenhänge erklären. In Abbildung 3.14 ist dargestellt, dass sich die signifikanten Kookkurrenzen des Wortes „radiation“ verändern.[10] Auch die Kookkurrenzen der Wörter im Umfeld von „radiation“ nehmen zu. Dies deutet darauf hin, dass das Wort und dessen Kontext am 15.03.2011 stärker in die Berichterstattung eingebettet ist und komplexer diskutiert wird. Diese Entwicklung ist allein durch einen explorativen Umgang mit den Kookkurrenzdaten zu erkennen.[11] Jedoch ist nicht ersichtlich, welche Bedeutungen und Propositionen hinter der Veränderung stecken, sodass durch die Daten ein besseres Verständnis für diese Kontextveränderung entwickelt werden kann. Es ist lediglich ein anderer Verwendungszusammenhang vorhanden, der suggeriert, dass bestimmte Terme eine neue Rolle innerhalb der Thematisierung und der Wortverwendung einnehmen.

Innerhalb der Betrachtung von Themen ist es jedoch bedeutend, welche Handlungen, Modifikationen oder Veränderungen die Themenwörter während der Themenentfaltung erfahren. Deshalb ist es wichtig, die Relationen mit einer Information zu versehen, welche die Position die Worte in den lokalen Kontexten zueinander einnehmen. Aus diesem Grund kann die Zählung der Kookkurrenzen auf nachbarschaftliches gemeinsames Auftreten beschränkt werden (Heyer, 2006, vgl. S. 31). Der lokale Kontext wird auf die jeweilige Nachbarschaft der betrachteten Wortform beschränkt. In Heyer (2006) bezieht sich der Nachbarschaftskontext auf die direkte Nachbarschaft der Wörter, die links und rechts von einem analysierten Wort zu finden sind. Mit dieser

[10] Die Visualisierungen sind im Leipzig Corpus Miner erstellt, der von Niekler u. a. (2014b) entwickelt wird.

[11] Die Exploration von Daten ist in Abschnitt 2.2.2 erläutert.

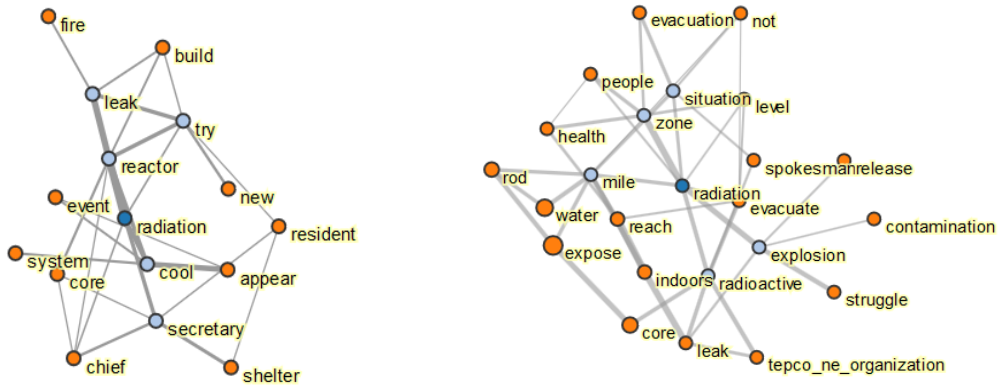

Abbildung 3.14: Beispiel der Entwicklung von Kookkurrenzen des Wortes „radiation“ am 11.03.2011 (linker Graph) und am 15.03.2011 (rechter Graph) in der Online-Ausgabe des Guardian.

Einschränkung lassen sich Qualifizierungen[12], Modifikationen[13] und Veränderungen [14] analysieren. Wird nur die direkte Nachbarschaft betrachtet, so können nicht alle semantischen Relationen, die in einer Kontexteinheit stehen, erfasst werden, da oft nicht signifikante Nachbarschaften durch sprachliche Mittel vorhanden sind. In dem Satz – „If its fuel rods reached too high a temperature, they would have melted at least partly.“ – würde die Relation (melted - partly) nicht zu erfassen sein. Für die interpretierbare Formulierung von Propositionen gilt, dass „[. . .]sich die Bestimmung nicht an den stilistischen und grammatischen Besonderheiten der jeweiligen Formulierung orientiert, sondern die zugrundeliegende Bedeutung erfasst, [. . .]“ (Früh, 2007, S. 247) wird. Für die weiteren Ausführungen wird der Begriff der Nachbarschaft deshalb auf einen Betrachtungshorizont erweitert, sodass nicht nur die direkte Nachbarschaft zählt, sondern die Nachbarschaften in einem bestimmt definierten Umfeld des Wortes. So können zu einer Wortform lokale linke Kontexte und lokale rechte Kontexte erfasst werden. Es werden alle Wortformen vor einer Wortform in einem Satz, Absatz oder Dokument als linke Nachbarn oder Vorgänger gezählt und alle Wortformen die folgen als rechte Nachbarn oder Nachfolger gezählt. Bei einem Betrachtungshorizont von nur einem Wort vereinfacht sich das Verfahren auf direkte Nachbarschaften. Diese Form der Anwendung bedarf einer Repräsentation der Dokumente als Folge natürlicher Zahlen $v_i = \{v_1, \ldots, v_n\}$, wie es auf Seite 45 gezeigt wird, sodass die Dokumente als

[12] Maßeinheiten

[13] Attribute von Objekten

[14] Handlungen von Objekten

Folge der Wortformen dargestellt werden. Nur so kann der linke und rechte Kontext der Wörter bei der Verarbeitung erhalten bleiben.

Um die Signifikanz einer Kookkurrenz zu bestimmen, werden mehrere Signifikantmaße vorgeschlagen und evaluiert (Dunning, 1993; Bordag, 2007, 2008; Heyer, 2006). Dunnings Log-Likelihood Statistik erweist sich als geeignet, um Kookkurrenzsignifikanzen für unterschiedliche Anwendungen zu berechnen. Dieser statistische Test weist asymptotische Eigenschaften aus und ist deshalb besser als andere Tests für Untersuchungen von Texteigenschaften geeignet (Dunning, 1993). Im Falle von zwei Miteinander auftretenden Ereignissen, ist der Test wie eine χ^2 Verteilung mit einem Freiheitsgrad definiert. Der wichtigste Nutzen ist, dass statistische Signifikanztests mit weniger Daten durchgeführt werden können und seltene und häufige Ereignisse bezüglich ihrer Signifikanz verglichen werden können (Dunning, 1993). Das Maß ist allerdings anfällig für sehr seltene Ereignisse und überschätzt deren Signifikanz, sodass die seltenen Ereignisse aus dem Signifikanztest ausgeschlossen werden müssen (Frequenz-Pruning) (Biemann, 2012, vgl. S. 48).

Für die Berechnung der gemeinsam auftretenden Wortformen, sei es als Nachbarschaftskookkurrenz oder in einem lokalen Kontext, können verschiedene Vorverarbeitungsschritte vorgenommen werden, um das Ergebnis zu beeinflussen. Mit dem Ziel, dass relevante Propositionen innerhalb der Thematisierungen erkannt werden sollen, muss vorher festgelegt werden, welche Bestandteile die Propositionen haben und wie sie extrahiert werden können. Da eine Proposition ein ausgedrückter Sachverhalt, in einem bestimmten Kontext ist (van Dijk, 1980; Brinker, 1988, vgl. S. 27, vgl. S. 24), können die Bestandteile, die zur Darstellung nötig sind, eingegrenzt werden. Innerhalb der empirischen Kategorienbildung wird untersucht, wie in den zugrunde liegenden Texten Akteure oder Objekte verschiedene Handlungen eingehen oder welche Eigenschaften sie besitzen, um daraus ein theoretisches Konstrukt als Kategorie abzuleiten. Besonders in Versuchen, die empirische Kategorienbildung zu formalisieren, wird auf die Arbeit mit Propositionen gesetzt, die ein Handlungs- oder Zustandskonzept beschreiben (Früh, 2001; Fillmore, 1968). Früh ergänzt, dass die Handlungs- und Zustandskonzepte durch Valenzstellen erweitert werden sollen, um die Propositionen darzustellen. Da Valenzen vor allem zu Verben formuliert werden und Objekt - Attribut - Handlung - Zusammenhänge darstellen (Ágel, 2000, vgl. S. 7), ist die Erweiterung von Früh nichts anderes als die Aussage, dass Handlungen, Zustände und Modifikationen durch Objekte im Text erweitert werden müssen, auf die sie sich beziehen. Ein solches Zustandskonzept muss von funktionalen oder stilistischen Spracheigenschaften befreit sein und unabhängig vom Text funktionieren, da es nicht

Teil der Syntax ist (Jackendoff, 1990, vgl. S. 46). In Hausser (2000, vgl. S. 67) werden Substantive, Verben, Adjektive und Adverbien als grundlegende Bestandteile einer Proposition genannt. Diese Sichtweise soll vor allem Orientierung im Text schaffen. Diskursive oder Argumentative Strukturen sind interessant, werden aber für eine thematische Betrachtung an dieser Stelle ausgeschlossen. Nach dieser Anforderung ist es demnach sinnvoll, nur Nomen, Verben und Adjektive in die Analyse der Nachbarschaftskookkurrenzen einfließen zu lassen. Für die empirische Kategorienbildung in der Inhaltsanalyse, die Früh (2007, 2001) vorschlägt, können so statistisch-syntagmatische Relationen, die zur Exploration und Ableitung der Proposition nach van Dijk (1980); Mackeldey (1987) geeignet sind, berechnet werden. Die Kookkurrenzen repräsentieren semantische Relationen wie Handlungen, Handlungsträger oder Eigenschaften (Heyer, 2006, vgl. S. 148).

In Abbildung 3.15 wird dargestellt, wie sich Propositionen oder Aussagen aus gerichteten Graphen explorativ erarbeiten lassen.[15] In der oberen Grafik existiert eine gerichtete Kante „radiation → level → normal“. Zu einem späteren Zeitpunkt verändert sich die Relation in „radiation → level → high“. Die Darstellung von gerichteten Kanten, mit einer Beschränkung auf bestimmte Wortarten, macht es möglich, die Aussagen und deren Veränderung innerhalb einer Thematisierung zu erfassen, zu verschiedenen Zeitpunkten zu analysieren und dadurch den propositionalen Gehalt eines Themas explorativ zu erarbeiten.

Um die Interpretation und Ableitung von Propositionen anhand der Kookkurrenzen innerhalb eines Themas möglich zu machen, müssen die Kookkurrenzen in eine geeignete Darstellung überführt werden. Nach van Dijk (1980, vgl. S. 166) ist das semantische Gedächtnis bzw. die konzeptualisierte Darstellung einer Kategorie die Übersetzung der Relationen zwischen den Begriffen in ein Netzwerk dieser Begriffe. Gleichermaßen kann festgestellt werden, dass die Wortverbindungen im Gedächtnis des Menschen nicht hierarchisch verteilt sind, sondern als Clusterstrukturen vorliegen (Heyer (2006, vgl. S. 179), Steyvers u. Tenenbaum (2005); Watts u. Strogatz (1998)). Die Darstellung der Kookkurrenzen lässt sich nach diesem Verständnis visualisieren. Das gemeinsame Auftreten zweier Wortformen kann als Kante zwischen zwei Knoten in einem Graph dargestellt werden. Die Interkonnektivität der Kookkurrenzen untereinander stellt einen Graph mit Small World Eigenschaften dar (Ferrer i Cancho u. Solé, 2001). Ähnliche Eigenschaften weisen Graphen auf, die auf der Grundlage

[15] Die Visualisierungen sind im Leipzig Corpus Miner erstellt, der von Niekler u. a. (2014b) entwickelt wird.

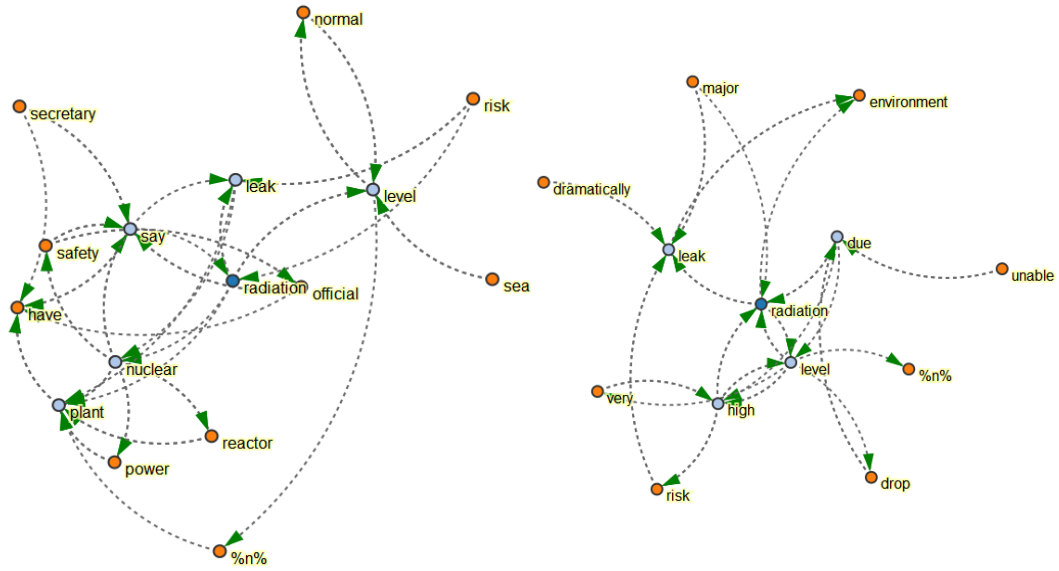

Abbildung 3.15: Beispiel der Entwicklung von gerichteten Kookkurrenzen des Wortes „radiation“ am 11.03.2011 (linker Graph) und am 15.03.2011 (rechter Graph) im Nachrichtenkorpus des Guardian.

sozialer Netzwerke gebildet werden (Heyer, 2006, vgl. S. 175). Solche Graphen lassen sich über unterschiedliche Verfahren visualisieren.

Zunächst muss gewährleistet werden, dass die Visualisierung von Graphen die Beziehung zwischen den Knoten erklärt (Fleischer u. Hirsch, 2001; Di Battista, 1999). Dabei können zwei Strategien der Visualisierungen unterschieden werden. Ein Ansatz benötigt eine vorherige Information über die Anordnung der Elemente in einem Graph. Wenn diese nicht bekannt ist wird mit einem anderen Visualisierungsansatz eine Anordnung selbstständig hergeleitet. Bei der Visualisierung sozialer Netzwerke, also auch anderen Graphen mit Small World Eigenschaften, gibt es keine Informationen über ein festes Raster oder ein optimales Layout der Elemente in der Graph-Visualisierung und die optimale Anordnung muss automatisch gefunden werden (Fleischer u. Hirsch, 2001, vgl. S. 216 ff.). In diesem Fall eigenen sich Methoden, die alle Knoten und Kanten in einem Graph als physikalische Komponenten mit interagierenden Kräften und minimaler Energie simulieren (Brandes (2001, vgl. S. 216 ff.), Di Battista (1999, vgl. S. 303 ff.)). In einem solchen System werden die Kräfte so definiert, dass der Ruhezustand eine möglichst optimale Visualisierung darstellt. Zu einem solchen Modell gehören immer ein physikalisches Modell und ein Algorithmus, der iterativ die anfangs zufällig angeordneten Elemente so platziert, das die Energie im Gesamtsystem minimiert wird. Es kommen Spring-Modelle, die eine Federspannung zwischen den Knoten eines Graphs simulieren, Gravitationsmodelle oder die Simulation von Ener-

giepotentialen zum Einsatz. Diese Gruppe von Visualisierungen wird Force Graphs genannt. Die in dieser Arbeit gezeigten Graphen basieren auf einem Gravitationsmodell, welches mit der Visualisierungsbibliothek D3 (Bostock u. a., 2011) umgesetzt wird.

Die Kookkurrenzanalyse innerhalb einer Thematisierung wird genutzt, um einzelne Wortverwendungszusammenhänge innerhalb von Thematisierungen zu erkennen. Durch die Analyse gerichteter Kookkurrenzen zwischen bestimmten Wortarten ist es möglich, Propositionen innerhalb eines Themas zu identifizieren. Neben der Darstellung eines Themas als Nominal- oder Wortgruppe (TDT, Topic-Modelle) oder als Fokus (Topic-Modelle) kann die Darstellung einzelner signifikanter Propositionen ergänzt werden. Diese Vorgehensweise hilft, auf Basis einer explorativen Strategie, Themenbeschreibungen und -erklärungen zu entwickeln. Die Erzeugung von Makropropositionen zur Kategorienbildung kann dadurch zwar nicht automatisiert, jedoch semi-automatisch unterstützt werden. Durch den Einsatz dieses Verfahrens ist es ferner möglich, die automatische Erfassung der Themen durch TDT und Topic-Modelle mit einer detaillierten Themenbeschreibung auf Grundlage von deren propositionalem Gehalt zu ergänzen.

3.4 Häufigkeiten, Messgrößen und Zeitreihen in Themen

Die Bestimmung quantitativer Größen, wie Häufigkeiten und Anteile der Themen, muss für die inhaltsanalytische Auswertung der automatischen Methoden möglich sein. Das betrifft die quantitativen Aussagen, die nach der Zuordnung von Textthemen innerhalb eines Korpus bestimmt werden können. Nur so lassen sich Hypothesen oder Fragestellungen mit konkreten und vergleichbaren Werten testen und beurteilen. Allein die Bestimmung der Themenmerkmale über einen gewissen Zeitraum und deren Vergleich ist eine wichtige Voraussetzung für die Bestimmung von Trends in der Themenanalyse (Merten, 1995, vgl. S. 150). Im Folgenden Abschnitt werden deshalb quantitative Messgrößen für die Auswertung der Ergebnisse automatischer Methoden in synchronen und diachronen Textquellen definiert.

3.4.1 Themenhäufigkeit

Die Bestimmung einer Größe, welche Aufmerksamkeit einem Thema zukommt, ist ein Indikator für den Nachrichtenwert eines Themas (Rössler, 2005, vgl. S. 225). Je mehr Nachrichtenfaktoren ein Thema erfüllt, desto wahrscheinlicher ist dessen Aufnahme in ein publiziertes Medium. Wenn demnach bestimmt wird, wie groß der

Anteil eines Themas an einem Korpus ist, so kann darauf geschlossen werden, welche Nachrichtenfaktoren und Nachrichtenbestandteile für eine hohe Aufmerksamkeit und eine hohes Nachrichtenaufkommen sorgen. Für die Messung der Themenanteile und -häufigkeiten können verschiedene Variablen bestimmt werden. Im Fall der Analyse von Nachrichtenartikeln müssen neben Artikelzählungen auch formale Kategorien, wie beispielsweise die Seitenzahl, die Flächenanteile von Bildern und der Textanteil bestimmt werden. Daraus ergibt sich der Anteil eines Themas an der Gesamtberichterstattung eines Nachrichtenmediums. In digitalen Textkollektionen ist das Wissen über die physische Beschaffenheit der Texte, wenn sie aus gedruckten Texten digitalisiert wurden oder als gedruckter Text existieren, allerdings meist nicht vorhanden. Demnach kann nur die Anzahl S_k und deren relativer Anteil P_k an der Textmenge, die einem Thema k zugeordnet ist, in der Textkollektion bestimmt werden. Es gibt Unterscheidungen in der Art der Berechnung, denn der Anteil an einem Thema k kann unterschiedlich zustande kommen. Folgende Größen können für ein Thema bestimmt werden:

1. Die Zählung der Dokumente D_k, die einem Thema k zugeordnet sind, ausgedrückt durch
$$S_k^D = |D_k| \,. \tag{3.9}$$

2. Zählung der Token $w_{i,d}$, die einem Thema k zugeordnet sind, ausgedrückt durch
$$S_k^W = \sum_d^M \sum_i^N c_{d,i},\ c_{d,i} = \begin{cases} 1 & \text{für } Thema(w_{i,d}) = k \\ 0 & \text{sonst} \end{cases},\ d \in D, \tag{3.10}$$
mit dem Term $w_{i,d}$, dem $i$ten Wort im $d$ten Dokument der Dokumentmenge D_k.

3. Der relative Anteil der Dokumente an einer Textkollektion,
$$P_k^D = S_k^D / |D|. \tag{3.11}$$

4. Der relative Anteil der Token,
$$P_k^W = S_k^W / V, \tag{3.12}$$
wobei V die Summe aller Token in der Textkollektion repräsentiert.

Wenn die Themenaufmerksamkeit bzw. Themenintensität über einen Zeitraum abgebildet werden soll, so muss die Bestimmung der Anzahl und des relativen Anteils für

unterschiedliche Zeitpunkte ermittelt werden. In die Berechnungen für die Zeitpunkte T fließen nur Dokumente und Token mit einem Bezug zu einzelnen Zeitpunkten t ein.

1. Zählung der Dokumente $D_{k,t}$, die einem Zeitpunkt t zugeordnet sind, ausgedrückt durch
$$S^D_{k,t} = |D_{k,t}|\,, \tag{3.13}$$
wobei $D_{k,t}$ die Dokumente beinhaltet, die zu einem Zeitpunkt t einem Thema k zugeordnet sind. Die Berechnung ist analog zu Formel 3.9.

2. Die Zählung der Token $S^W_{k,t}$ zu einem Zeitpunkt t wird berechnet, indem die Dokumentmenge für die Formel 3.10 auf die Menge D_t beschränkt wird,
$$S^W_k = \sum_d^M \sum_i^N c_{d,i},\ c_{d,i} = \begin{cases} 1 & \text{für } Thema(w_{i,d}) = k \\ 0 & \text{sonst} \end{cases},\ d \in D_t. \tag{3.14}$$

3. Relativer Anteil der Dokumente an einer Textkollektion zu einem Zeitpunkt t, ausgedrückt durch
$$P^D_{k,t} = S^D_{k,t}/|D_t|. \tag{3.15}$$

4. Relativer Anteil der Token,
$$P^W_{k,t} = S^W_{k,t}/|V_t|, \tag{3.16}$$
wobei V_t die Summe aller Token in der Dokumentmenge D_t repräsentiert.

Die Darstellung der normierten, relativen Anteile an einer Textkollektion spielen bei der Erstellung von Zeitreihen eine wichtige Rolle. Durch die unterschiedlichen Dokumentmengen in einzelnen Zeiträumen können Trends und Anteile durch absolute Zählungen nicht abgebildet werden. Die Schwankungen innerhalb einer Zeitreihe können durch unterschiedliche Dokumentmengen verursacht sein und die Aussagekraft geht verloren.

3.4.2 Worthäufigkeit

Innerhalb eines Themas kann bestimmt werden, welche der Wörter $w_{n,k}$ in einem Thema mit der absoluten Häufigkeit $S^W_{n,k}$, bzw. der relativen Häufigkeit $P^W_{n,k}$, vorkommen. Es können folgende Größen ermittelt werden:

1. Die Zählung der Wörter w_n die einem Thema k zugeordnet sind, ausgedrückt durch

$$S^W_{n,k} = \sum_d^M \sum_i^N c_{d,i,n},\ c_{d,i,n} = \begin{cases} 1 & \text{für } Thema(w_{i,d}) = k \text{ und } w_{i,d} = n \\ 0 & \text{sonst} \end{cases} \quad d \in D, \tag{3.17}$$

   wobei n einen Term (Type) aus dem Vokabular darstellt, der innerhalb des Themas beobachtet werden soll.

2. Der relative Anteil der Wörter $w_{n,t}$ an einem Thema,

$$P^W_{n,k} = S^W_{n,k}/V_k, \tag{3.18}$$

   wobei V_k die Anzahl aller dem Thema k zugeordneten Token darstellt.

Werden die Anteile der Wörter über die Zeit dargestellt, kann die Variation der Häufigkeit einzelner Terme in einem Thema sichtbar gemacht werden. Eine Analyse, welche Wörter innerhalb eines Themas zu verschiedenen Zeitpunkten genutzt werden, ist dadurch möglich. Gewissermaßen kann bestimmt werden, welchem speziellen Interesse, Akteursspektrum oder welcher kommunikativen Absicht ein Thema unterliegt. Die Bestimmung der auf Zeitpunkte orientierten Größen erfolgt durch folgende Berechnungen.

1. Die Zählung der Wörter $w_{n,t}$, die einem Thema k zu einem Zeitpunkt t zugeordnet sind, ausgedrückt durch

$$S^W_{n,k,t} = \sum_d^M \sum_i^N c_{d,i,n},\ c_{d,i,n} = \begin{cases} 1 & \text{für } Thema(w_{i,d}) = k \text{ und } w_{i,d} = n \\ 0 & \text{sonst} \end{cases},\ d \in D_t, \tag{3.19}$$

   wobei n einen Term (Type) aus dem Vokabular darstellt, der innerhalb des Themas beobachtet werden soll.

2. Der relative Anteil der Wörter $w_{n,t}$ an einem Thema,

$$P^W_{n,k,t} = S^W_{n,k,t}/V_{k,t}, \tag{3.20}$$

   wobei $V_{k,t}$ die Anzahl aller dem Thema k zum Zeitpunkt t zugeordneten Token darstellt.

Im Fall der Topic-Modelle können einem Dokument mehrere Themen zugeordnet sein. Themen können unterschiedlich stark repräsentiert sein, was durch die Verteilung $p(\mathbf{z}|d)$ ausgedrückt wird. Die Themen, welche die höchste Wahrscheinlichkeit in

dieser Verteilung aufweisen, prägen den Inhalt eines Dokuments. Mit einem Schwellwert kann bestimmt werden, welche Themen relevant für ein Dokument sind. Einem Thema werden demnach nur Dokumente zugeordnet, die dieses Thema mit einem vorgegebenen Anteil enthalten. Da ein Dokument durch mehrere Themen repräsentiert wird, ist der Anteil der Wörter, die sich im gleichen Dokument befinden, aber einem anderen Thema zugeordnet sind, interessant. Diese Wörter geben zusätzliche Kommunikationsabsichten oder Aspekte wieder. Zusätzlich zu den Größen $S^W_{n,k,t}$ und $P^W_{n,k,t}$ bzw. $S^W_{n,k}$ und $P^W_{n,k}$ kann deshalb auch gemessen werden, wie sich eine Wortform n innerhalb der Dokumente verhält, die einem Thema k zugeordnet sind. Das Wort wird nicht nur gezählt, wenn es einem Thema direkt zugeordnet ist. Es wird immer gezählt wenn es in einem Dokument vorkommt, welches das Thema k zu einem Anteil enthält. Aus diesem Grund wird die Fallunterscheidung aus 3.17 und 3.19 ersetzt.

$$\begin{cases} 1 & \text{für } w_{i,d} = n \\ 0 & \text{sonst} \end{cases}, \; d \in D_k \tag{3.21}$$

Dies gilt auch für den Fall verschiedener Zeitpunkte t.

$$\begin{cases} 1 & \text{für } w_{i,d} = n \\ 0 & \text{sonst} \end{cases}, \; d \in D_{k,t}, \tag{3.22}$$

Die Dokumentmengen D_k und $D_{k,t}$ sind die mit einem Thema k assoziierten Dokumente. Die Größen V_k und $V_{k,t}$ müssen in diesem Fall entsprechend durch die Zählung der Token, welche in D_k und $D_{k,t}$ erhalten sind, ersetzt werden.

Die Dokumentfrequenz einzelner Terme kann interessant sein. So können alle Dokumente gezählt werden, die eine Wortform n enthalten. Die Berechnung der Dokumente, welche zu einem Thema k zugeordnet sind und einen Term n enthalten ist die Mächtigkeit der Menge

$$S^D_{k,n} = |D_{k,n}| \,. \tag{3.23}$$

Für den Fall mehrerer Zeitpunkte ergibt sich die Menge

$$S^D_{k,n,t} = |D_{k,n,t}| \,. \tag{3.24}$$

Um auch hier relative Größen berechnen zu können, wird S^D jeweils mit der Dokumentmenge D_k und $D_{k,t}$ normiert.

$$P^D_{k,n} = S^D_{k,n} / |D_k| \tag{3.25}$$

$$P^D_{k,n,t} = S^D_{k,n,t}/|D_{k,t}| \tag{3.26}$$

Es ist vorstellbar, dass Gruppen von Wörtern festgelegt werden, die für eine bestimmte Annahme, Hypothese oder ein Konzept stehen. Die Häufigkeiten können für die betreffenden Wortformen akkumuliert werden. Weiterhin ist es möglich, dass bestimmte Wortformen, wie zum Beispiel Namen, Orte oder Institutionen, addiert werden, sodass globale Eigennamenanteile in den Themen nachgewiesen werden können. Es ist vorstellbar, dass beispielsweise Personennamen in Kategorien aufgeteilt werden. So können Personen, die innerhalb einer Thematisierung genannt werden, einem Spektrum von wissenschaftlichen oder politischen Akteuren zugeordnet werden. Die Phasen einer Thematisierung können so, anhand der Akteursstruktur, quantitativ beschrieben werden, wie es in (Kolb, 2005) vorgeschlagen wird.

3.5 Zusammenfassung

In diesem Kapitel werden zwei grundsätzliche Methoden für die automatische Themenbestimmung in digitalen Textkollektionen dargestellt und untersucht. Die Verfahren unterscheiden sich durch die Art und Weise, wie die Themen repräsentiert werden. Das gezeigte TDT-Verfahren stellt die Themen als gemittelten Termvektor für alle Themendokumente dar, während die Topic-Modelle die Themen in Form einer Wahrscheinlichkeitsverteilung abstrahieren. Die Topic-Modelle können mehrere Themen für ein Dokument vergeben, während das Clusterverfahren hinter TDT dies nicht vorsieht. In beiden Fällen ist die Repräsentation der Themen über Termvektoren oder Wahrscheinlichkeitsverteilungen geeignet, um eine Darstellung von Wortgruppen, im Sinne der kategorialen Abgrenzung, zu realisieren. Um ein höheres Verständnis der Themen zu gewährleisten, muss diese Form der Themendarstellung erweitert werden. Durch Einbeziehung von Kookkurrenzanalysen kann die Repräsentation durch Bedeutungsstrukturen und den propositionalen Gehalt der Themen erweitert werden. Um die Themenstrukturen quantitativ bewerten zu können, müssen die Wörter und Dokumente zu den Themen in der Textkollektionen zugeordnet werden. Durch die Größen in Abschnitt 3.4, können aus den Zuordnungen der Themen Zeitreihen extrahiert werden, die es erlauben, die Themenanteile an einer Gesamtkollektion zu beurteilen. Weiterhin ist es möglich, Strukturen und Bestandteile innerhalb der Themen zu analysieren. Neben einer ergänzenden Kookkurrenzanalyse erlaubt dieses Vorgehen die explorative Analyse und Identifikation der Phasen und Schlüsselereignisse innerhalb einer Thematisierung.

Aus den vorgeschlagenen Verfahren lassen sich so Themenzusammenfassungen, Zeitreihen und Kookkurrenzgraphen erstellen. Diese dienen als Input für eine explorative Analyse der Themenstruktur eines Korpus. Über die Zeitreihen und Kookkurrenzgraphen können quantitative Vergleiche und die Prüfung von Hypothesen realisiert werden. Die Durchführung explorativer Analysen mit den dargestellten Verfahren wird im nächsten Kapitel exemplarisch durchgeführt und durch Möglichkeiten der visuellen Darstellung ergänzt. Die Leistungsfähigkeit und die Validität der Methoden wird evaluiert, um deren Gültigkeit für inhaltsanalytische Aufgaben nachzuweisen.

Kapitel 4

Exemplarische Analyse

Um die Anwendbarkeit der Verfahren, die in Kapitel 3 besprochen werden, für die praktische Arbeit zu demonstrieren, wird im Folgenden eine exemplarische Themenanalyse besprochen. Die Anwendung orientiert sich an einer inhaltsanalytischen Fragestellung, deren Zweck die Analyse einer konkreten Thematisierung ist. Es sollen drei verschiedene Tageszeitungen im Zeitraum vom 1. März 2011 bis zum 30. April 2011 untersucht werden. Darunter sind die „Süddeutsche Zeitung“ (SZ) und die „tageszeitung“ (TAZ), welche in deutscher Sprache erscheinen. Die dritte Quelle besteht aus den Artikeln der Online-Ausgabe des Guardian, welche in englischer Sprache veröffentlicht werden. Jede Tageszeitung stellt einen Einzelkorpus dar. Dies soll die Vergleichbarkeit der Ergebnisse, die aus unterschiedlichen Textquellen und Korpora erzeugt werden können, zeigen. Jedes Korpus wird in Tages-Zeitscheiben repräsentiert, sodass jeder Artikel einem Tag zugeordnet werden kann. Diese Information wird aus den Metadaten der Textquellen entnommen. Die Zusammensetzung der Quellen ist in Tabelle 4.1 dargestellt.

	W	V	D	$\varnothing D_t$	$\varnothing W_t$	Speicher
Süddeutsche Zeitung (SZ)	4.904.625	261.009	9.621	158	80.403	32 MB
Guardian	11.897.822	244.139	20.927	343	195.046	79 MB
die Tageszeitung (TAZ)	2.950.557	183.133	8.253	135	48.369	20 MB
Süddeutsche Zeitung (SZ) (Grundform)	4.928.953	210.573	9.621	158	80403	31 MB
die Tageszeitung (TAZ) (Grundform)	2.960.220	148.509	8.253	135	48.369	18 MB

Tabelle 4.1: Darstellung der zu analysierenden Korpora. Die Größe D bzw. D_t bezeichnen die Dokumentanzahl für den Korpus und dessen durchschnittliches Dokumentaufkommen an einem Tag dar. Mit V wird die Anzahl aller Types bezeichnet und W bzw. W_t bezeichnet die Anzahl der Token im Korpus.

Die Analysen sollen die praktische Relevanz für die Inhaltsanalyse zeigen. Im Rahmen der Untersuchen werden zwei Themen definiert, die analysiert werden sollen. Dies sind die Thematisierungen

- Fukushima, die Thematisierung der Ereignisse nach dem Erdbeben und dem darauf folgendem Tsunami am 11.3.2011 in Japan (JAP) und
- die Ereignisse des Bürgerkriegs in Libyen (LIB),

die in den Korpora analysiert werden sollen. Die Themen sollen hinsichtlich der inhaltlichen Struktur und der Themeneigenschaften untersucht werden. Die Analyse der inhaltliche Struktur besteht aus der Abstraktion des Themengehalts, die den Inhalt eines Themas gemäß der linguistischen Thementheorien aus Abschnitt 2.1.3 wiedergibt. Die Themeneigenschaften sollen vor allem die Maße aufgreifen, welche in Abschnitt 3.4 definiert werden. So sollen Akteursstrukturen, Trends, Schlüsselereignisse oder Nachrichtenfaktoren der ausgewählten Themen sichtbar werden. Für die Bearbeitung der Analyse sind die Schritte

- explorative Identifikation relevanter Themenstrukturen,
- Bestimmung der relevanten Dokumente,
- Bestimmung der Themeninhalte,
- Bestimmung der Themeneigenschaften,
- und Bildung geeigneter Zeitreihen notwendig.

Diese Schritte folgen dem unter Abschnitt 2.2.2 vorgestellten Ablauf einer explorativen Analyse in zwei Phasen. In einer ersten Phase, dem explorativen Browsing, werden die relevanten Themenzusammenhänge identifiziert. In einer zweiten Phase, der fokussierten Suche, werden die identifizierten Themen genutzt, um Dokumentmengen für die Detailanalyse eine Themas einzugrenzen. Es werden qualitative und quantitative Zugriffe über Zusammenfassungen, Wortanalysen oder Zeitreihen hergestellt, die eine detaillierte Analyse der Themen zulassen. Auch die Evaluierung der Reliabilität und der Validität der Verfahren kann in der zweiten Phase durchgeführt werden, indem die Zeitreihen und Themenzusammenfassungen genutzt werden.

Für die Analyse wird ein Programmpaket verwendet, welches die Verfahren und die nötigen Verarbeitungsschritte implementiert. Das Paket wurde in Java erstellt und wird im Anhang A ausführlicher beschrieben. Für die explorative Analyse wurde eine grafische Oberfläche entwickelt, die es erlaubt, nötige Informationen über die

thematische Struktur eines Textkorpus zu visualisieren und Themen zu identifizieren (Niekler u. Jähnichen, 2012). Die Programme arbeiten auf der Basis von Zeitscheiben. Es ist mit der verwendeten Software aber auch möglich, einen Textkorpus ohne Beachtung der Zeitscheibenstruktur zu berechnen.

4.1 Vorbereitung und Verarbeitung

Die Korpora müssen für die Anwendung der Verfahren in eine maschinenlesbare Struktur übersetzt werden. Hierfür werden die Texte in jeder Zeitscheibe zunächst transformiert und jedes Dokument als Termvektor repräsentiert. Für die Transformation werden die Dokumente folgendermaßen vorverarbeitet:

- Die Zeichen werden in Kleinbuchstaben(lowercase) überführt.
- Es werden Absätze identifiziert und markiert.
- Leere Zeilen oder Absätze in den Dokumenten werden gelöscht.
- Es wird eine Eigennamenerkennung durchgeführt und die entsprechenden Namen werden in den Dokumenten annotiert. Es werden Grundformen der Wörter erzeugt (Lemmatisierung), um die Auswirkung dieser Vorverarbeitung bei Themenanalysen beurteilen zu können. Die Grundform ist, anders als der Wortstamm, die morphologisch unflektierte Form eines Wortes im Text.

Die Eigennamenerkennung wird jeweils mit einem Modell für das Deutsche und für das Englische durchgeführt. Hierfür wird der CRF Klassifizierer aus dem StanfordNLP Paket verwendet (Finkel u. a., 2005).

Die Dokumente werden tokenisiert und in jeweils einen Dokumentvektor V überführt. Für die Erstellung der Dokumentvektoren wird eine Wortliste generiert, in der jeder Type auf einer fortlaufenden Nummer (ID) abgebildet wird. Gleichzeitig werden Wortformen, die sich nur in Groß- und Kleinschreibung unterscheiden auf die gleiche ID abgebildet (lowercase-Transformation). Durch diese Vorverarbeitung wird der Effekt der Groß- und Kleinschreibung eliminiert. Eigennamen können im Verlauf der Analyse identifiziert und selektiert werden. Weiterhin werden die Korpora für die SZ und die TAZ dupliziert und noch einmal einer Grundformreduktion unterzogen. Mit diesem zusätzlichen Schritt soll der Nutzen der Morphologieerkennung für die automatische Themenidentifikation untersucht werden.

Die Zeitscheibenstruktur wird auf der Basis von Tagen erstellt, sodass für jeden Tag eine Zeitscheibe mit den jeweiligen Dokumenten existiert. Für jedes der fünf

Korpora ergibt sich demnach eine Struktur von 61 Tageszeitscheiben.[1] Mit dieser Struktur ist es möglich, tageweise Batches, Topic-Modelle und Zeitreihen zu bilden.

4.2 Bestimmung relevanter Themen

Die Durchführung der ersten Phase einer explorativen Suche besteht aus dem explorativen Browsen. In den folgenden Abschnitten wird dieser Schritt mit Hilfe von Textdateien und einer geeigneten grafischen Oberfläche demonstriert. Die zwei Möglichkeiten sollen die Identifikation relevanter Themenstrukturen unterstützen und so die Festlegung einer Dokumentmenge für die Weiterverarbeitung erlauben. In diesem Schritt ist es theoretisch möglich, die Modellqualität zu beurteilen, Parameter anzupassen und die Modelle in einer modifizierten Form neu zu berechnen. So kann der Abstraktionsgrad der Themen bis zu einer gewünschten Auflösung angepasst werden. Für die folgenden Betrachtungen werden für die explorative Arbeit und die detaillierte Auswertung aber festgelegte Modelle berechnet und evaluiert, um eine detaillierte nachvollziehbare Untersuchung dieses Schritts angeben zu können. Für jedes der fünf Korpora wird einmal für alle Zeitscheiben

- ein Clustering über TDT,
- ein Topic-Modell mit LDA und
- ein Topic-Modell mit HDP erstellt.

Jede Tageszeitscheibe wird als Stapel an den jeweiligen Algorithmus übergeben. Für jede Dokumentmenge innerhalb einer Tageszeitscheibe werden jeweils ein LDA Topic-Modell und ein HDP-LDA Topic-Modell berechnet, sodass zwei Modelle für eine Tageszeitscheibe existieren. Zusätzlich werden je vier „globale“ Modelle (LDA , HDP-LDA) für die kompletten Korpora berechnet. Für die globalen Modelle wird je ein Modell mit hohem Abstraktionsgrad und ein Modell mit sehr detaillierten Themen berechnet, sodass pro Verfahren zwei Modelle erstellt werden. Damit soll im Verlauf der praktischen Anwendung der Einfluss der Granularität dargestellt werden. Die Topic-Modelle in den Tageszeitscheiben werden jeweils genutzt, um eine sequenzielle Themenbestimmung, wie in Niekler u. Jähnichen (2012) und Abschnitt 3.2.4 dargestellt wird, durchzuführen. Für das TDT-Verfahren wird für jedes Korpus je eine Berechnung durchgeführt. Für die Berechnung der Modelle werden jeweils folgende Verarbeitungs- und Transformationsschritte gewählt:

[1] Eine detaillierte Darstellung wird in Tabelle 4.1 auf Seite 111 gezeigt.

- **Pruning:** Für die Verfahren werden jeweils nur Wörter einbezogen, die mindestens 3 mal pro Zeitscheibe verwendet werden. So wird die Berechnung effizienter und Datenpunkte, die keine statistische Bedeutung haben stören die Verfahren nicht. Für die „globalen“ Topic-Modelle wird das Pruning auf den Gesamtkorpus angewendet, sodass nur Wörter genutzt werden, die mindestens 3 mal im Gesamtkorpus zu finden sind.

- **Stopwortfilterung:** Für alle Verfahren werden die häufigsten Wörter für die jeweilige Sprache entfernt, da diese keine thematische Bedeutung tragen und syntaktische Eigenschaften der Sprache repräsentieren.

Bei Topic-Modellen lohnt es sich, die Leistungsfähigkeit in Bezug auf die mögliche Themenauflösung bzw. deren Granularität zu evaluieren. Aus diesem Grund werden für die „globalen“ Themen zwei Varianten für jedes Korpus und jedes Topic-Modell berechnet, deren Themenanzahl differiert. Für die zeitscheibenbasierte Berechnung wird darauf allerdings verzichtet, da in einer Zeitscheibe zahlenmäßig weniger Themen zu erwarten sind. Dort ergibt sich die Heterogenität automatisch, indem die wenigen Einzelthemen in den Zeitscheiben miteinander kombiniert werden. Es ist nicht hilfreich, sehr grobe Themen in den Zeitscheiben zu wählen, in der Hoffnung, dass diese für wichtige Themen über die Zeitscheiben hinweg stabil aussehen. Dies ist bei den kleinen Dokumentmengen in den Zeitscheiben nicht zu erwarten. Um die unterschiedlichen Topic-Modelle zu berechnen werden deshalb folgende Festlegungen getroffen:

- In den Zeitscheiben werden jeweils Modelle für HDP-LDA und LDA berechnet. Für den Parameter der Themenanzahl, im LDA Verfahren, wird $K = 60$ gewählt. Die Hyperparameter α und β werden auf ein Optimum geschätzt, welches von der Themenanzahl abhängt.[2] Dies ist bei diesem Verfahren sehr praktisch, da die Festlegung die Hyperparameter für eine Analyseperspektive keine Relevanz mehr haben. Die HDP-LDA Modelle schätzen die Anzahl der Gruppen bzw. Themen selbst. Zur Steuerung der Anzahl der Themen dient ein Hyperparameter, der die Spezifizität der Term-Topic Verteilungen $p(\mathbf{w}|z)$ beeinflusst. Je nachdem, welche Form dieser Verteilung durch die Hyperparameter vorgegeben wird, müssen mehr oder weniger Themen durch das Verfahren produziert werden, um das Modell optimal unter anzupassen. Im HDP-LDA Modell kann der Parameter für die Beeinflussung der Verteilung $p(\mathbf{w}|z)$ auch

[2] Die Vorgehensweise wird auf S. 87 beschrieben.

als β bezeichnet werden. In der Abbildung 3.7 auf Seite 86 wird gezeigt, dass der Einfluss auf die Themenanzahl durch diesen Parameter im Intervall 0.01 - 0.3 am größten ist. Angelehnt an diese Betrachtung wird festgelegt, dass mit einem Parameterwert von $\beta = 0.2, \alpha = 0.1$ ca. 20 - 60 Themen zu erwarten sind.

- Die globalen Modelle, die mit der LDA berechnet werden, sind pro Korpus in jeweils 60 und 250 Themen unterteilt. Die Hyperparameter α und β werden geschätzt.

- Die globalen Modelle, die mit dem HDP Ansatz erstellt werden, müssen über die Hyperparameter konfiguriert werden, sodass das Modell in der Anzahl der Themen variabel ist. Die Anzahl der Themen hängt von der Spezifizität der Verteilung $p(\mathbf{w}|z)$ ab, die über den Parameter β gesteuert werden kann. In der Zeitscheiben-Variante wird mit einem Parameterwert von 0.2 gearbeitet. Dieser Wert wird in den globalen Modellen für die Berechnung von wenigen, grob aufgelösten Themen eingestellt. Das Ergebnis soll mit dem aus dem LDA Verfahren vergleichbar sein, sodass ca. 60 und 250 Themen gefunden werden sollen. In der Abbildung ist dieser Bereich zwischen 0.02 (ca. 250 Themen) und 0.2 (ca. 30 Themen) zu finden. Um die Unterschiede und die Nutzbarkeit unterschiedlicher Themengranularitäten zu testen, werden diese zwei Extreme getestet und der Wert von β für die Berechnung vieler detaillierter Themen auf 0.02 gesetzt.

Die Ergebnisse lassen sich für jede Zeitscheibe oder für den Gesamtkorpus anhand von Wortvektoren visualisieren. Die zu explorierende Datenmenge beträgt an dieser Stelle K Themen pro gerechnetem Verfahren. Aus dieser Menge müssen nun jeweils die relevanten Themen für eine Analyse identifiziert werden. Dazu müssen die Wortvektoren der Themen dargestellt und durch eine qualitative Betrachtung eines Nutzers beurteilt werden. Da sehr prominente Themen oft sehr heterogen oder in mehreren eng verwandten Themen diskutiert werden, können mehrere Wortvektoren für ein Thema entstehen. Zur Visualisierung der thematischen Wortvektoren kann auf einfache Textdateien oder grafische Oberflächen zurückgegriffen werden. Grafische Oberflächen bieten sich vor allem an, wenn die unterschiedliche Gewichtung der Worte, innerhalb einer Thematisierung, zur Erhöhung der Übersicht herangezogen werden soll. Diese können z.B. genutzt werden, um die Größe der Wortdarstellung zu beeinflussen und so den Fokus auf wichtige Wortformen einer Thematisierung lenken zu können. Die Arbeit mit WordClouds ist weit verbreitet für diesen Zweck (Lohmann

u. a., 2009). Die Arbeit an den für die Analysen verwendeten Korpora soll an dieser Stelle mit Text-Dateien und grafischen Oberflächen skizziert werden. Das Ergebnis der explorativen Analyse zur Identifikation relevanter Themen in den Daten, wird mit Textdateien produziert und am Ende dieses Abschnitts präsentiert.

4.2.1 Explorative Analyse mit Textdateien

Die Suche nach relevanten Themen erfolgt mit dieser Arbeitsweise über die Erstellung einer Textdatei. Soweit möglich, sollten die Themen einen chronologischen Bezug behalten. Dennoch unterscheiden sich die Darstellungen der Wortvektoren der Themen je nach Verfahren und Verarbeitung der Daten. Für die Analyse werden drei unterschiedliche Darstellungen unterschieden. Kurze Beispiele für die jeweiligen Darstellungsmöglichkeiten werden in Anhang B gegeben.

1. **TDT:** Das Verfahren arbeitet mit Dokumentstapeln (Batches), die nach und nach an das Verfahren übergeben werden. Das Ergebnis sind Cluster von Dokumenten und deren Termvektoren. Da das Verfahren mit jedem Dokumentstapel die Dokumentfrequenz aller bekannten Types aufgrund der bereits gesehenen Dokumente neu berechnet, muss das Verfahren nur auf das Zeitscheibenformat angewendet werden. Ohne die Aktualisierung der Dokumentfrequenzen wird die Entwicklung von Themen nicht modelliert und das Verfahren wird fehlerhaft. Das schließt eine globale Berechnung mit diesem Verfahren aus. Aus den Termvektoren der Dokumente, die einem Cluster zugeordnet sind, kann ein Mittelwertvektor berechnet werden, der für die Darstellung eines Themas genutzt wird. Die resultierende Textdatei stellt die Mittelwertvektoren zeilenweise dar. Die Reihenfolge der Themen orientiert sich am Datum des erstmaligen Auftretens eines Themenclusters.

2. **Topic-Modell (Zeitscheiben):** Bei dieser Vorgehensweise wird innerhalb jeder Zeitscheibe ein Topic-Modell mit K Themen erstellt. Diese werden mit einem Vergleich über die Zeitscheiben hinweg verbunden.[3] Für die Darstellung werden die verbundenen Themen gruppiert. Für jede Zeitscheibe existieren somit Darstellungen einer jeden Themenverteilung $p(\mathbf{w}|z)$. Themen, die gruppiert sind, werden chronologisch aneinandergereiht. Die Sortierung in der gesamten Textdatei ist so organisiert, dass eine Themengruppe und deren erstmaliges

[3] Eine detaillierte Prozessbeschreibung befindet sich in Abschnitt 2

Auftreten chronologisch in die Textdatei eingefügt wird. Zu jedem Thema wird das Datum der zugehörigen Zeitscheibe mit in der Textdatei vermerkt.

3. **Topic-Modell (global):** Bei dieser Berechnungsart ist es nicht möglich, ein Thema einem konkreten Zeitstempel zuzuordnen, da alle Dokumente in die Berechnung einfließen. Deshalb können die Wortvektoren für die Themen lediglich zeilenweise in die Textdatei eingetragen werden. Die spätere chronologische Zuordnung der Themen muss nachträglich über die zugeordneten Dokumente geschehen.

Jedes Thema innerhalb der Textdateien bekommt einen Identifikator, um eine eindeutige Zuordnung innerhalb der Datenstrukturen zu ermöglichen. Für die Verknüpfung zweier Themen sequenzieller Zeitscheiben mit Algorithmus 2 (siehe Anhang D S. 211) muss eine Mindestähnlichkeit für die Kosinusdistanz definiert werden.[4] Um das Verhalten der Themenverknüpfung zu evaluieren wird die Themenzuordnung mehrfach mit verschiedenen Schwellwerten wiederholt. Die Anwendung der Kosinusdistanz basiert auf den je 10 wahrscheinlichsten Termen aus den Themen, die in Abbildung 3.12 auf S. 94 gezeigt werden. Für den Schwellwert von $s(A, B)$ werden, für die Bildung der zeitscheibenbasierten Themengruppen, die Werte 0.3, 0.4, 0.5 und 0.6 erprobt. Zum Vergleich werden globale Topic-Modelle berechnet. Für jedes Korpus und die Verfahren HDP-LDA und LDA wird je ein Topic-Modell mit hoher Themenanzahl und je ein Modell mit niedriger Themenanzahl berechnet. Durch die unterschiedlichen Themenzahlen soll überprüft werden, wie sich die Ergebnisse der zeitscheibenbasierten und der globalen Modelle unterscheiden. Die unterschiedliche Granularität der globalen Themen soll demonstrieren, wie sich Themen in Haupt- und Nebenthemen aufspalten und zusammenfassen. Die genaue Unterscheidung zwischen Neben- und Hauptthemen wird später erläutert.[5] Aus den Verfahren (LDA, TDT, HDP-LDA), ob global oder zeitscheibenbasiert, und deren Parametrisierungen (Hyperparameter, Schwellwerte) resultieren unterschiedliche Themenmengen für jedes Korpus. Eine Übersicht ist in Tabelle 4.2 dargestellt. Für die Mindestähnlichkeit im clusterbasierten TDT-Verfahren, wird der als optimal vorgeschlagene Wert von 0.21 verwendet (Allan u. a., 2005).[6] In Abschnitt 3.1 wird dargestellt, dass die Verwendung unterschiedlicher Schwellwerte zu viele Fehler und Unsicherheiten verursacht. Deshalb wird auf jedem Korpus genau eine TDT-Berechnung erstellt.

[4] Die Kosinusdistanz ist in Formel 3.8 definiert.

[5] Die entsprechende Passage ist auf S. 121 zu finden.

[6] Die Prozedur wird in Algorithmus 1 auf S. 1 gezeigt.

Die Tabelle 4.3 zeigt für jedes Korpus die Aufschlüsselung aller angewendeter Verfahren und Parameter. Für die analysierten Themen LIB und JAP werden je zwei Spalten für jede Berechnung dargestellt. Eine repräsentiert die Anzahl der manuell identifizierten Hauptthemen (K_H), die andere repräsentiert die identifizierten Nebenthemen (K_N). Die Berechnungen werden jeweils durch den Namen des verwendeten Verfahrens gekennzeichnet. Die jeweilige Parametrisierung für den Ähnlichkeitsschwellwert in den zeitscheibenbasierten Topic-Modellen ist tiefgestellt gekennzeichnet. Die globalen Topic-Modelle sind mit tiefgestellten $\check{G}$, für die Modelle mit wenigen Themen, und mit $\widehat{G}$, für Modelle mit vielen Themen, gekennzeichnet.

Für die zeitscheibenbasierten Topic-Modelle muss die Granularität der Themen vorgegeben werden werden.[7] Obwohl die Granularität konstant für jede Zeitscheibe bleibt, ist die Bildung der Themenketten über die Zeitscheiben hinweg dennoch von einem Schwellwert abhängig. Die Definition eines Schwellwertes für $s(A, B)$ aus der Berechnung 3.8 führt zu verschiedenen Themenmengen für unterschiedliche Schwellwerte. Bei hohen Schwellwerten entstehen mehr Themen, da weniger Themen einander zugeordnet werden. Es wird aufwendiger, die relevanten Themen unter den Themenketten zu identifizieren. Es stellt sich aber die Frage, welche Fehler bei der Verkettung der Themen durch den Algorithmus gemacht werden, und wie sich Haupt- und Nebenthemen trennen lassen, wenn die Themen mit unterschiedlichen Schwellwerten verkettet werden. Bei den globalen Topic-Modellen resultiert nur die Anwendung des HDP-Modells in unterschiedlichen Themenzahlen, da hier nicht, wie bei der LDA, der Parameter K direkt beeinflusst wird. Für die Auswertung der resultierenden Themen müssen die relevanten Termvektoren bzgl. der gesuchten Thematisierungen JAP und LIB identifiziert werden. Die Termvektoren werden als kommagetrennte Wortlisten zeilenweise und nach der oben dargestellten Ordnung in eine Textdatei geschrieben.[8] Es ist sinnvoll, dass bei den zeitscheibenbasierten Topic-Modellen alle Themen, die zusammengefasst sind, gruppiert in die Datei geschrieben werden.

Um einen Blick auf die grundsätzlich relevanten Themen zu werfen, wird jede Zeile in der Textdatei markiert, deren Wortzusammensetzung einer bestimmten Erwartung an die gesuchten Themen entspricht. Dies wird mit einem regulären Ausdruck für jedes der Themen realisiert. Es wird gewissermaßen nach Zeilen gesucht, die bestimmte Schlüsselwörter enthalten. Diese werden markiert und manuell nach ihrer Relevanz

[7] Für die Anwendung der zeitscheibenbasierten LDA wird $K = 50$ gewählt. Die Parameter α und β werden jeweils geschätzt. Für die Berechnung der HDP wird $\beta = 0.2, \alpha = 0.1$ gewählt. Mit dieser Einstellung werden jeweils 30-60 pro Zeitscheibe berechnet.

[8] Beispielhaft ist in Anhang B dargestellt, wie die Textdateien ausgegeben werden können.

Modell	K	Modell	K
Süddeutsche Zeitung (SZ)			
$\text{LDA}_{\widehat{G}}$	250	$\text{LDA}_{\check{G}}$	60
$\text{HDP}_{\widehat{G}}$	200	$\text{HDP}_{\check{G}}$	20
$\text{LDA}_{0.3}$	989	$\text{LDA}_{0.5}$	1851
$\text{LDA}_{0.4}$	1471	$\text{LDA}_{0.6}$	2128
$\text{HDP}_{0.3}$	1517	$\text{HDP}_{0.5}$	2406
$\text{HDP}_{0.4}$	2036	$\text{HDP}_{0.6}$	2630
TDT	2125		
Guardian			
$\text{LDA}_{\widehat{G}}$	250	$\text{LDA}_{\check{G}}$	60
$\text{HDP}_{\widehat{G}}$	211	$\text{HDP}_{\check{G}}$	41
$\text{LDA}_{0.3}$	1046	$\text{LDA}_{0.5}$	1784
$\text{LDA}_{0.4}$	1287	$\text{LDA}_{0.6}$	2205
$\text{HDP}_{0.3}$	1691	$\text{HDP}_{0.5}$	2276
$\text{HDP}_{0.4}$	1875	$\text{HDP}_{0.6}$	2797
TDT	2482		
die Tageszeitung (TAZ)			
$\text{LDA}_{\widehat{G}}$	250	$\text{LDA}_{\check{G}}$	60
$\text{HDP}_{\widehat{G}}$	172	$\text{HDP}_{\check{G}}$	46
$\text{LDA}_{0.3}$	1064	$\text{LDA}_{0.5}$	1943
$\text{LDA}_{0.4}$	1552	$\text{LDA}_{0.6}$	2252
$\text{HDP}_{0.3}$	1272	$\text{HDP}_{0.5}$	2099
$\text{HDP}_{0.4}$	1745	$\text{HDP}_{0.6}$	2343
TDT	1073		
Süddeutsche Zeitung (SZ) (Grundform)			
$\text{LDA}_{\widehat{G}}$	250	$\text{LDA}_{\check{G}}$	60
$\text{HDP}_{\widehat{G}}$	203	$\text{HDP}_{\check{G}}$	36
$\text{LDA}_{0.3}$	935	$\text{LDA}_{0.5}$	1820
$\text{LDA}_{0.4}$	1395	$\text{LDA}_{0.6}$	2111
$\text{HDP}_{0.3}$	1347	$\text{HDP}_{0.5}$	2147
$\text{HDP}_{0.4}$	1817	$\text{HDP}_{0.6}$	2396
TDT	1764		
die Tageszeitung (TAZ) (Grundform)			
$\text{LDA}_{\widehat{G}}$	250	$\text{LDA}_{\check{G}}$	60
$\text{HDP}_{\widehat{G}}$	154	$\text{HDP}_{\check{G}}$	32
$\text{LDA}_{0.3}$	1014	$\text{LDA}_{0.5}$	1963
$\text{LDA}_{0.4}$	1566	$\text{LDA}_{0.6}$	2238
$\text{HDP}_{0.3}$	1295	$\text{HDP}_{0.5}$	2203
$\text{HDP}_{0.4}$	1845	$\text{HDP}_{0.6}$	2444
TDT	1038		

Tabelle 4.2: Verteilung der Themen in unterschiedlichen Verfahren und Parametrisierungen der Themenähnlichkeit.

beurteilt. Dabei fällt auf, dass durch unterschiedliche Ausprägungen und Facetten, mehrere nicht durch den technischen Prozess verbundene Termvektoren zu einer Thematisierung gehören können. Dies unterstreicht, dass die Verfahren die manuelle qualitative Kontrolle nicht überflüssig machen. Sie helfen aber enorm, thematische Strukturen schnell zu identifizieren und zu messen. Die regulären Ausdrücke werden wie folgt verwendet.

- **Regular Expression für JAP:**
  `^.*\"(japan|erdbeben|fuksushima|wasser).*\"$` und
  `^.*\"(japan|earthquake|fuksushima|water).*\"$`
- **Regular Expression für LIB:**
  `^.*\"(libyen|gaddafi|rebell).*\"$` und
  `^.*\"(libyia|gaddafi|rebel).*\"$`

Die Relevanzbewertung der Themen wird sehr pragmatisch durchgeführt. Eine Relevanz ergibt sich, wenn innerhalb eines Themas ein Suchwort aus dem jeweiligem regulären Ausdruck zu finden ist und der Kontext, der im Termvektor zu erkennen ist, Wörter enthält, die dem Thema zuzuordnen sind. So wäre ein Thema, dessen Wortvektor sowohl „japan“ als „naturkatastrophe“ enthält, für das Thema JAP relevant. Mancher Kontext ist allerdings nicht sofort zuordenbar. So existieren für die Thematisierung JAP Termvektoren für Themen, die eines der Suchworte enthalten, aber im Kontext mit anderen Wörtern wie „endlager“„energiewende“ oder „reaktorsicherheit“ enthalten. Dies ist ein Hinweis, dass die hauptsächliche Thematisierung von Nebenthemen begleitet wird und die Themenkontexte sehr heterogen sein können. Je nachdem, ob der eigentliche Konflikt oder die Ereignisse besprochen werden oder ein durch die Ereignisse ausgelöster oder beeinflusster Diskurs in den Medien stattfindet. Da die beeinflussten Themen, die dennoch im Kontext der angesprochenen Thematisierungen auftreten können, eine Evaluierung verfälschen können, wird bei der Identifikation darauf geachtet, diese von den Hauptthemen abzutrennen. Die Nebenthemen sollen nicht ignoriert werden, sondern es muss ebenso evaluiert werden, in welcher Weise sie die Thematisierung und die möglichen Analysen beeinflussen. Auszüge aus den paradigmatischen Wörtern der Themen- und der Nebenthemen sind in der folgenden Liste dargestellt. Die Übersicht soll klar machen, dass es, unabhängig vom Verfahren, immer mehrere Aspekte einer Thematisierung geben kann. Die Identifikation relevanter Aspekte muss, abhängig von der Fragestellung, beim Anwender liegen.

- **JAP:**

  - Hauptthema: japan, menschen, erdbeben, katastophe, tepco, fukushima, reaktoren
  - Atomkraftwerke/Reaktoren/Kernkraft allgemein: netz, reaktor, krümmel, strom, alt, akw, generation, siedewasserreaktoren, betrieb
  - Strahlung und dessen Folgen: reaktor, jod, plutonium, radioaktiven, krebs, dosis, gefahr, millisievert
  - Energiepolitik/Atomausstieg/erneuerbare Energie: moratorium, ausstieg, ausbau, erneuerbaren, energien, stromleitungen, atomausstieg
  - Tschernobil: kinder, atomkraft, frauen, mutter, tschernobyl, symbol, männer, folgen, liquidatoren, unfall
  - Witschaftliche Folgen: yen, länder, problem, wirtschaft, toyota
  - Protest: atomkraftgegner, kampagne, geplant, proteste, rufen, umgebung, zahlreichen, katastrophe
  - Internationale Atomenergie-Organisation (IAEO, engl. International Atomic Energy Agency, IAEA): iaeo, who, gesundheitlichen, fragen, fukushima, katastrophe, wissenschaftler

- **LIB:**

  - Hauptthema: libyen, armee, gaddafi, tripolis, aufständischen, mittelmeer, regime, milizen, al-gaddafi, staatschef, waffen
  - Flüchtlinge/Lampedusa: insel, lampedusa, hafen, boot, flüchtlinge, europa, libyen, menschen, grenze
  - Politik/Sicherheitsrat/UN: libyen, nato, gaddafi, un-resolution, enthaltung, flugverbotszone, außenpolitik, eu, sicherheitsrat
  - Ölpreis: sanktionen, libyen, ölkonzerne, probleme, suche, phase, barrel, bp, kampf
  - Syrien/Arabischer Frühling: bahrain, demonstranten, libyen, land, welt, demokratiebewegung, regierung, ägypten, syrien, türkei, reformen

Eine Zerlegung in Haupt- und Nebenthemen gelingt nur, wenn die Themenmodelle in einem ausreichendem Detailgrad arbeiten. Die Beeinflussung ist bei Topic-Modellen einfach möglich, da durch die Wahl der Themenanzahl oder der Hyperparameter eine direkte Beeinflussung möglich ist. Die Veränderung des Detailgrades bei Clustermethoden, beispielsweise TDT, führt jedoch schnell zu Fehlern. Die optimale Auflösung muss für eine Analyseaufgabe getestet werden. Das Interesse des Analysten steht im

Vordergrund. Wird eine zusammenfassende Darstellung von Themen erstellt, wie z.B. die Trennung der Texte in Ressorts, so kann die Auflösung sehr grob sein. In Fällen, wo eine nachträgliche Trennung der Details einer Thematisierung nötig ist, müssen mehr Themen identifiziert werden. Eine feste Aufteilung in Themen existiert für einen Korpus nicht. Die Analysestrategie legt fest, welche Aussagekraft und Details untersuchte Thematisierungen aufweisen müssen.

Die oben angerissenen Wortkontexte helfen, die Inhalte der relevanten Themen abzugrenzen. Die Kontexte müssen allerdings nicht unbedingt vorher definiert sein, da sich die Interpretation während der explorativen Analyse ergibt. Es ist möglich, ein vorher deduktiv festgelegtes Verständnis einer Thematisierung für die Relevanzbewertung festzulegen. Dies ist analog zur deduktiven Definition und Beschreibung einer Kategorie in der Inhaltsanalyse anzusehen. Die im Folgenden beschriebe Relevanzbewertung wird selbstständig durchgeführt und objektiv an den beschriebenen Vorgaben der Wortbestandteile durchgeführt. Die Zeilen in den Textdateien werden mit regulären Ausdrücken markiert und anhand des Kontextes eines Themenvektors bewertet.

Aus der Relevanzbewertung resultiert für jedes Korpus und jedes Verfahren eine Menge an identifizierten Wortvektoren. Die Vektoren werden dem Haupthema und den Nebenthemen zugeordnet. In Tabelle 4.3 sind alle als relevant empfundene Themen quantitativ erfasst. In den folgenden Ausführungen wird anhand der zwei Tabellen in diesem Abschnitt geklärt, welche Themengranularität zu ausreichend detaillierten Ergebnissen führt und welcher Arbeitsaufwand jeweils nötig ist. Darauf aufbauend wird geklärt, welche Parametrisierung und Granularität für unterschiedliche Aufgaben sinnvoll und handhabbar ist.

Bei der Berechnung einer Menge von Themen mit einem der Verfahren lassen sich anhand der Tabellen folgende Beobachtungen treffen. Die Anzahl der Themen lässt sich bei der Anwendung von Topic-Modellen durch die Festlegung der Modellparameter bestimmen. Werden die Topic-Modelle in mehreren Zeitscheiben angewendet, so hängt die Anzahl der Themen davon ab, welche Themen durch den festgelegten Schwellwert miteinander verkettet werden. Je mehr Themen gefunden werden, desto mehr relevante Themen und Nebenthemen lassen sich identifizieren. Dies spielt vor allem bei den zeitscheibenbasierten Themenverkettungen eine Rolle. Hier werden pro Zeitscheibe ca. 50 Themen berechnet. Im Vergleich zu den globalen Topic-Modellen wird die Gesamtmenge aller Dokumente dadurch schon in viel mehr Themen unterteilt. Insgesamt werden, wenn auf jeder der 60 Zeitscheiben eine LDA mit 60 Themen berechnet wird, 3600 Themen bestimmt, die im Nachhinein über deren Wortbestand-

Modell	K^H_{JAP}	K^H_{JAP}	K^H_{LIB}	K^N_{LIB}	Modell	K^H_{JAP}	K^N_{JAP}	K^H_{LIB}	K^N_{LIB}
Süddeutsche Zeitung (SZ)									
$\text{LDA}_{\widehat{G}}$	2	3	2	1	$\text{LDA}_{\check{G}}$	1	1	1	1
$\text{HDP}_{\widehat{G}}$	1	2	1	2	$\text{HDP}_{\check{G}}$	1	0	1	0
$\text{LDA}_{0.3}$	8	9	4	4	$\text{LDA}_{0.5}$	19	42	13	13
$\text{LDA}_{0.4}$	15	35	8	2	$\text{LDA}_{0.6}$	22	50	–	–
$\text{HDP}_{0.3}$	10	10	6	2	$\text{HDP}_{0.5}$	11	46	16	8
$\text{HDP}_{0.4}$	11	26	4	4	$\text{HDP}_{0.6}$	18	56	–	–
TDT	4	15	7	0					
Guardian									
$\text{LDA}_{\widehat{G}}$	1	3	2	4	$\text{LDA}_{\check{G}}$	1	1	1	0
$\text{HDP}_{\widehat{G}}$	1	1	1	1	$\text{HDP}_{\check{G}}$	1	1	1	1
$\text{LDA}_{0.3}$	2	0	4	1	$\text{LDA}_{0.5}$	4	5	12	11
$\text{LDA}_{0.4}$	3	4	3	3	$\text{LDA}_{0.6}$	8	11	–	–
$\text{HDP}_{0.3}$	4	1	2	2	$\text{HDP}_{0.5}$	5	3	7	3
$\text{HDP}_{0.4}$	4	1	8	1	$\text{HDP}_{0.6}$	8	14	–	–
TDT	1	0	1	6					
die Tageszeitung (TAZ)									
$\text{LDA}_{\widehat{G}}$	1	5	2	1	$\text{LDA}_{\check{G}}$	1	3	1	1
$\text{HDP}_{\widehat{G}}$	1	5	1	1	$\text{HDP}_{\check{G}}$	1	2	1	1
$\text{LDA}_{0.3}$	6	31	4	4	$\text{LDA}_{0.5}$	29	70	30	12
$\text{LDA}_{0.4}$	23	39	16	6	$\text{LDA}_{0.6}$	32	81	–	–
$\text{HDP}_{0.3}$	15	21	8	1	$\text{HDP}_{0.5}$	18	57	28	3
$\text{HDP}_{0.4}$	18	45	14	4	$\text{HDP}_{0.6}$	35	78	–	–
TDT	6	10	5	3					
Süddeutsche Zeitung (SZ) (Grundform)									
$\text{LDA}_{\widehat{G}}$	2	3	1	2	$\text{LDA}_{\check{G}}$	1	2	1	1
$\text{HDP}_{\widehat{G}}$	1	3	1	1	$\text{HDP}_{\check{G}}$	1	1	1	1
$\text{LDA}_{0.3}$	6	15	5	4	$\text{LDA}_{0.5}$	16	53	6	17
$\text{LDA}_{0.4}$	10	32	5	7	$\text{LDA}_{0.6}$	27	60	–	–
$\text{HDP}_{0.3}$	5	11	4	4	$\text{HDP}_{0.5}$	12	34	6	20
$\text{HDP}_{0.4}$	7	20	5	7	$\text{HDP}_{0.6}$	21	37	–	–
TDT	3	8	3	4					
die Tageszeitung (TAZ) (Grundform)									
$\text{LDA}_{\widehat{G}}$	2	4	2	1	$\text{LDA}_{\check{G}}$	1	3	1	1
$\text{HDP}_{\widehat{G}}$	1	4	1	1	$\text{HDP}_{\check{G}}$	1	1	1	0
$\text{LDA}_{0.3}$	9	21	8	1	$\text{LDA}_{0.5}$	19	65	24	8
$\text{LDA}_{0.4}$	14	45	17	2	$\text{LDA}_{0.6}$	34	79	–	–
$\text{HDP}_{0.3}$	2	11	10	2	$\text{HDP}_{0.5}$	13	42	16	19
$\text{HDP}_{0.4}$	9	27	16	10	$\text{HDP}_{0.6}$	24	70	–	–
TDT	5	7	4	4					

Tabelle 4.3: Verteilung der relevanten Themen in unterschiedlichen Verfahren.

teile und mittels der Kosinusdistanz verkettet werden. Je höher der Schwellwert für die Mindestähnlichkeit zweier Themen gesetzt wird, desto mehr isolierte Themen bleiben nach diesem Prozess übrig. Die Bildung von einzelnen Modellen für jede Zeitscheibe führt zu Resultaten, deren Themen an die kleinen Mengen angepasst sind. Die Wortzusammensetzungen der Themen können zwischen den Zeitscheiben sehr heterogen sein. Gleichermaßen relevante Themen sind dennoch nicht über einen Wortvergleich zu verknüpfen. Dies passiert beispielsweise, wenn sich die Thematisierung zwischen den Tagen mit unterschiedlichen Einzelaspekten beschäftigt und nur wenige Dokumente zu einem Thema verfasst werden. In diesem Fall kann eine Relevanz oder Zusammengehörigkeit nur durch manuelles Validieren bestimmt werden.

Im Gegensatz zu den kleineren Zeitscheibenmodellen werden bei den globalen Modellen G mit vergleichbaren Modellparametern nicht vergleichbare Anzahlen von Themen generiert. Obwohl zwei verschiedene Modellgranularitäten erzeugt werden, entstehen weitaus weniger Themen als bei der Verkettung. Die Anzahl der als relevant empfundenen Themen ist geringer. Topic-Modelle bilden auf großen Datenmengen generellere, allgemeingültige Wortverteilungen in den Topics als in kleineren Mengen.

Das clusterbasierte Verfahren TDT erzeugt mit einem fest vergebenen Schwellwert von 0.21 für die Clusterähnlichkeit sehr viele Themen. Da das Verfahren den direkten Vergleich dokumentbasierter Termvektoren zur Grundlage hat, werden die latenten Semantiken der Wortverteilungen in den Dokumenten nicht beachtet. Die Dokumente, die einer Gruppierung zugeordnet werden, enthalten demnach übereinstimmende Wörter. Das zeigt, dass die gefundenen Themen sehr ereignisbezogen sind und mehr Einzelgeschichten, anstatt komplexe Themen, repräsentieren. Der Abstraktionsgrad ist gering und kann nicht problemlos angepasst werden, wie in Kapitel 3 gezeigt wird.

Anhand der identifizierten Themenvektoren und deren Verknüpfung kann Phase zwei der explorativen Analyse erfolgen. Über die festgelegten Dokumentmengen werden Analysen von Themen- und Worthäufigkeiten durchgeführt. Auch die Analyse von Aussagen innerhalb eines Themas ist mit einer fokussierten Dokumentmenge möglich. Diese Untersuchungen sind in Abschnitt 4.3, 4.4 und 4.5 dokumentiert.

4.2.2 Explorative Analyse mit grafischen Oberflächen

Im letzten Abschnitt wird der Umgang mit Ergebnissen und die Identifikation relevanter Themen, auf der Grundlage von Textdateien, diskutiert. Aus der darin beschriebenen Vorgehensweise und aus den Darstellungsmöglichkeiten der Themenwörter können Anforderungen für die Umsetzung grafischer Oberflächen zur explorativen Analyse von Themenstrukturen abgeleitet werden.

- Die Wortvektoren der Themen, $\mathbf{w}_{avg,t}$ für TDT und $p(\mathbf{w}|z)$ für die Topic-Modelle, können als Auflistung von Wörtern dargestellt werden. Die Reihenfolge der Terme kann durch deren Gewichtung im Wortvektor oder durch die Wortfrequenz des jeweiligen Terms im Korpus erfolgen. Die Schriftgröße eines dargestellten Wortes kann variieren, sodass darüber eine zusätzliche visuelle Information für ein Wort dargestellt werden kann. Je nachdem, ob die Sortierung sich nach der Wort-Gewichtung oder der Wortfrequenz im Korpus richtet, kann die jeweils andere Information durch die Schriftgröße dargestellt werden. Eine weit verbreitete Technik, um eine Menge an Wörtern auf diese Weise darzustellen, sind Tag Clouds (Lohmann u. a., 2009).

- In einer Vorverarbeitung für die Rohtexte können Named Entities im Text markiert werden. Diese können innerhalb der Wortvektoren farblich markiert werden, falls sie ein relevanter Bestandteil der gewichteten Wortvektoren sind.

- Die Sortierung der Themen kann anhand der zugeordneten Dokumente oder Wörter erfolgen. Beim TDT-Verfahren kann einfach gezählt werden, wie viele Dokumente einem Cluster zugeordnet werden. Bei Topic-Modellen gibt es mehrere Möglichkeiten eine Sortierung vorzunehmen. Die Verteilungen der Themen innerhalb eines Dokuments kann herangezogen werden, um zu bestimmen, welche Dokumente einem Thema zugeordnet sind. Wird beispielsweise angenommen, dass ein Dokument mit einem Thema assoziiert wird, wenn es einen Mindestanteil von 10, 20 oder 30 Prozent an diesem Dokument hat, ergeben sich für jedes Thema Zählungen von Dokumenten. Diese Häufigkeiten für jedes Thema können zur Sortierung der Themen in einer grafischen Darstellung herangezogen werden. Weiterhin kann die Verteilung $p(\mathbf{z})$ innerhalb des Korpus verwendet werden. Jedes Thema kann gemäß der Wahrscheinlichkeit in eine Reihenfolge gebracht werden.

Diese Darstellungskonzepte können durch unterschiedliche Annahmen erweitert werden. Bei einem Topic-Modell können beispielsweise nur die Dokumente einem Thema zugeordnet werden, bei denen es die größte Wahrscheinlichkeit im Dokument besitzt. So akzentuiert die resultierende Sortierung Themen, die eine hohe Wahrscheinlichkeit in vielen Dokumenten haben. Im Gegensatz zu Themen, die mit niedriger Wahrscheinlichkeit über sehr viele Dokumente verteilt sind, sind diese Themen semantischer und sind leichter zu interpretieren (Evans, 2014). Themen die in vielen Dokumenten mit niedriger Wahrscheinlichkeit auftauchen enthalten mehr Domänenwörter oder Stopwörter und repräsentieren syntaktische oder allgemeine Eigenschaften eines Korpus.

Abbildung 4.1: Termvektor aus der Thematisierung JAP vom 15.03.2011.

Abbildung 4.2: Termvektor aus der Thematisierung JAP vom 16.03.2011.

Abbildung 4.3: Termvektor aus der Thematisierung JAP für ein Nebenthema (Moratorium) vom 17.03.2011.

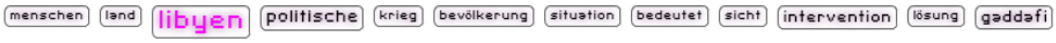

Abbildung 4.4: Termvektor aus der Thematisierung LIB vom 26.03.2011.

Diese Art der Sortierung setzt allerdings voraus, dass einem Dokument mehrere unterschiedliche Themen zugeordnet werden können. Für cluserbasierte Verfahren wie TDT kann sie nicht eingesetzt werden. Die beispielhafte Darstellung der Visualisierungsmöglichkeiten wird mit zwei Werkzeugen realisiert, die für die Vergleichbarkeit und Anwendung verschiedener Verfahren für die Inhaltsanalyse großer Korpora entwickelt wurden (Niekler u. a., 2012, 2014b). Die Beispiele in Abbildung 4.1, 4.2, 4.3 und 4.4 zeigen Wortvektoren, die für die Themen in einem Korpus oder in einer Zeitscheibe visualisiert wurden. Nennung von Orten, Personen oder Organisationen werden mit einer automatisch Namenserkennung annotiert und farblich markiert. In diesem Fall wird nur ein Ausschnitt aller Terme im Wortvektor visualisiert. In den gezeigten Beispielen liegt die Berechnung einer LDA zugrunde. Die jeweils wahrscheinlichsten 10 - 20 Terme für ein Topic werden aufsteigend nach der Wortfrequenz sortiert. Die Schriftgröße richtet sich nach der Wahrscheinlichkeit im Thema.

Werden diese Reihen von Wörtern untereinander in einer sortierten Liste dargestellt, kann ein schneller thematischer Überblick generiert werden. In Abbildung 4.5 ist diese Möglichkeit demonstriert. Für die Darstellung der Wortvektoren können mehr Terme herangezogen werden, sodass ein detaillierterer Überblick zu einem Thema hergestellt wird. Die Arbeit mit Textdateien für die Identifikation relevanter Themen

Abbildung 4.5: Darstellung mehrerer Wortvektoren als Liste, basierend auf Daten vom 24.03.2011.

Abbildung 4.6: Komplexe Darstellung eines Wortvektors als Tag Cloud. Die Daten basieren auf der Berichterstattung vom 15.03.2011.

bildet eine erste Annäherung an die Auswertung der Ergebnisse automatischer Themenanalysen. In diesen Auflistungen können Wortvektoren gesucht werden, die eine bestimmte Annahme einer relevanten Thematisierung wiedergeben. Die markierten Themen können gezielt nach Relevanz überprüft werden. Das Konzept kann auf grafische Oberflächen übertragen werden, sodass potentiell relevante Themen durch die Eingabe mehrerer Suchwörter visuell hervorgehoben werden. Durch das nachträgliche Verifizieren markierter Themen können, im Gegensatz zu reinen Textdateien, direkt Verknüpfungen im System mitgeführt werden. Das aufwändige Sammeln relevanter Bestandteile kann durch eine benutzerfreundliche grafische Eingabemöglichkeit unterstützt werden, indem zu verknüpfende Themen direkt in einer interaktiven Oberfläche gruppiert werden können. In Abschnitt 4.2.1 wird angesprochen, dass die Verkettung der Themen aus zeitscheibenbasierten Topic-Modellen fehlerhaft sein kann, wenn der Schwellwert für die Verkettung zu niedrig angesetzt wird. Bei einem hohen Schwellwert werden die zusammen gehörenden Themen mitunter nicht verkettet. In einer grafischen Oberfläche ist es problemlos möglich Fehlzuordnungen in einer Themenkette zu erkennen und zu markieren. Solche Markierungen werden, im Gegensatz zur Arbeit mit reinen Textdateien, bei der Auswertung einer Themenkette berücksichtigt. Die manuelle Bewertung der Themen wird so effizienter und zielorientiert und die Wei-

Abbildung 4.7: Termvektoren der Thematisierung JAP vom 15.03.2011 basierend auf dem Zeitungskorpus der TAZ.

Abbildung 4.8: Termvektoren der Thematisierung JAP vom 15.03.2011 basierend auf dem Zeitungskorpus der SZ.

terverarbeitung der Ergebnisse, wie die Selektion themenzugehöriger Dokumente und die Erstellung von Längsschnitten bzw. Zeitreihen, kann weitestgehend automatisiert geschehen.

Die einheitliche Darstellung der Wortvektoren dient zum qualitativen und explorativen Vergleich der Themen in unterschiedlichen Korpora. Zum einen ist es möglich, ähnliche Themen visuell zu erfassen. Zum Anderen können Themen nach Gemeinsamkeiten, Unterschieden oder nach unterschiedlichen Aspekten erfasst werden. Ein kleines Beispiel soll die Möglichkeiten verdeutlichen. In Abbildung 4.7 und 4.8 werden 2 visualisierte Wortvektoren dargestellt. Die Grundlage dieser Abbildungen sind jeweils LDA Berechnungen aus den Tageszeitscheiben vom 15.03.2011 der Korpora SZ und TAZ. Durch die teilweise Übereinstimmung, der in beiden Themen verwendeten Wörter, ist zu erkennen, dass eine ähnliche Thematik mit ähnlichen Aspekten angesprochen wird. Darauf weist die Übereinstimmung der Wörter *Reaktor*, *Reaktoren*, *Explosion*, *Meerwasser*, *Block* und *Kernschmelze* hin. Weiterhin sind aber auch Unterschiede zu erkennen. So werden im Thema, dass aus der TAZ berechnet wird, die Wörter *Brennstäbe*, *Wasser*, *Strahlung*, *Kühlung* und *Experten* verwendet. Im Thema aus der SZ werden die Wörter *Situation*, *Betreiber*, *Unfall*, *Techniker* und *Brennelemente* genannt. Durch die unterschiedliche Darstellung der einzelnen Wörter, aufgrund der Wahrscheinlichkeit oder Gewichtung im Wortvektor, können bei übereinstimmenden Wörtern dennoch die Unterschiede in den Themen sichtbar gemacht werden. Während die Thematisierung in der TAZ *Reaktoren*, *Brennstäbe*, *Strahlung*, *Block* und *Wasser* betont, stehen bei der SZ *Reaktoren*, *Fukushima*, *Kernschmelze*, *Block*, *Meerwasser* und *Explosion* im Vordergrund. Es ist abschließend also festzustellen, dass der Einsatz grafischer Oberflächen hilft, mit automatisch berechnetet Themenmodellen umzugehen, zu interagieren und um diese nachträglich zu validieren bzw. zu beurteilen.

4.2.3 Evaluation der explorativen Themenselektion

Validität der Themenverkettung

Die Messung der Themenstrukturen in unterschiedlichen Korpora ist von der Frage bestimmt, ob die gemessenen Zusammenhänge der Erwartung entsprechen und gesuchte Themenzusammenhänge abbilden. Diese Anforderung kann durch verschiedene Fehler in der Messung gestört werden. Bei der Verkettung der Themen durch den Vergleich der Wortbestandteile ist es möglich, dass Themen verbunden werden, die zwar Wörter teilen, aber dennoch nicht das gleiche Thema beschreiben. Ein zu hoch gewählter Schwellwert lässt hingegen eine gültige Verkettungen eventuell nicht zu. Anhand der Daten wird dieses Verhalten untersucht. Die Verkettung wird mit Schwellwerten von 0.3, 0.4, 0.5 und 0.6 durchgeführt und es werden die entstandenen relevanten Themenketten identifiziert. Es ist festzustellen, dass bei den Schwellwerten 0.3 und 0.4 Themen verkettet werden, sodass eine Mischung aus Haupt- und Nebenthemen entsteht. In einigen Fällen werden Themen verkettet, die keinen Zusammenhang zu einem relevanten Thema einer Zeitscheibe darstellen. An dieser Stelle sollen zwei kleine Beispiele zeigen, wie sich die Fehler bei der Verkettung äußern.

- Fehler bei der Zuordnung
  2011-04-04: rebellen, gaddafis, front, stadt, truppen, paar, kämpfen, brega, kilometer, männer, straßensperre
  2011-04-06: worden, menschen, demonstranten, land, opposition, demonstration, protesten, sicherheitskräfte
  2011-04-15: dienstag, straße, gesperrt, straßen, montag, entfernt, donnerstag, washington, mittwoch, hauptstadt, usa, like, all, getroffen,

- Vermischung von Themen und Nebenthemen
  2011-03-12: japan, erdbeben, tsunami, japanischen, beben, reaktoren, menschen, kilometer, abgeschaltet, tokio, land, batterien, katastrophe
  2011-03-14: japan, tschernobyl, deutschland, atomkraft, katastrophe, röttgen, deutschen, japanischen, restrisiko, umweltminister, wind, debatte, kernschmelze
  2011-03-15: japan, atomkraftwerke, baden-württemberg, atomkraft, neckarwestheim, cdu, taz, laufzeitverlängerung, sicherheit, angela_merkel

Es lassen sich zwei Fehler beschreiben. Zum einen kann es durch zu geringe Schwellwerte zu klaren Fehlzuordnungen der Themen kommen. Andererseits können die

Nebenthemen mit einem Hauptthema bzw. dem Thema, welches einem Ereignis zugeordnet ist, vermischt werden. Ab einem Schwellwert von 0.5 sind diese Fehler nicht mehr auffällig. Aber je höher der Schwellwert gesetzt wird, desto mehr Themengruppen entstehen, die ein Analyst prüfen muss, um deren Relevanz festzustellen. Allerdings kann mit höheren Schwellwerten genauer zwischen Haupt- und Nebenthemen unterscheiden werden. Eine hohe Genauigkeit bzw. Auflösung in mehrere Themenstränge hilft demnach, die einzelnen Aspekte einer Thematisierung zu erkennen. Je mehr Themen oder Aspekt in einer Thematisierung zusammengefasst werden, desto mehr Dokumente können einem Thema zugeschrieben werden. Dies verändert die Form einer Zeitreihe, welche die diachrone Zugehörigkeit von Dokumenten zu einer Thematisierung beschreibt. Ob diese Veränderung die Aussagekraft eines Längsschnittes verändert, muss in einer Auswertung der Zeitreihen geklärt werden. In Abschnitt 4.3 wird eine solche Evaluierung durchgeführt.

Reliabilität in unterschiedlichen Korpora

Die Anwendung der Verfahren auf unterschiedliche Korpora führt zu unterschiedlichen Ergebnissen. Um festzustellen, wie stabil die einzelnen Verfahren in unterschiedlichen aber themengleichen Textquellen arbeiten, werden die Verfahren auf die verschiedenen Korpora angewendet. Die Anzahl der gefundenen Themen wird bei jedem Durchlauf für jedes Korpus bestimmt, wie in Tabelle 4.2 gezeigt wird. In diesen unterschiedlichen Themenmengen können, wie in Tabelle 4.3 gezeigt, unterschiedliche, relevante Themen identifiziert werden. Die relevanten Themen, die in den unterschiedlichen Korpora durch ein Verfahren gefunden werden, müssen vergleichbar bleiben. Um dies zu testen, wird der Variationskoeffizient bestimmt, welcher die Varianz der Themenanzahl relevanter Themen zwischen den Korpora für ein Verfahren misst und normiert. Bei der Verkettung der Themen aus den Topic-Modellen ist der Variationskoeffizient $VarK$ für die jeweiligen Korpora und Verfahren zwischen 3% und 12%. Das Verfahren arbeitet relativ stabil in unterschiedlichen Korpora. Das Verfahren TDT bietet für alle Korpora nur einen Variationskoeffizient von 37%. Eine starke Abweichung der Themenanzahl ist beim TAZ-Korpus zu bemerken. Weiterhin ist der Unterschied zwischen der Themenanzahl im SZ-Korpus und dessen grundformreduzierter Variante auffällig. Oft weisen TF/IDF-basierte Verfahren eine starke Beeinflussung von der Dokumentlänge auf (Singhal u. a., 1996). Das TDT-Verfahren ist ebenso von der Dokumentlänge abhängig, wie andere TF/IDF basierte Verfahren. Im Vergleich der durchschnittlichen Dokumentlänge in den Korpora – Guardian 3214 Wörter, SZ

Modell	$VarK$
$LDA_{0.3}$	0.05
$LDA_{0.4}$	0.08
$LDA_{0.5}$	0.12
$LDA_{0.6}$	0.06
$HDP_{0.3}$	0.04
$HDP_{0.4}$	0.03
$HDP_{0.5}$	0.05
$HDP_{0.6}$	0.07
TDT	0.37
$HDP_{\check{G}}$	0.28
$HDP_{\widehat{G}}$	0.13

Tabelle 4.4: Varianzkoeffizient einzelner Verfahren, die auf die verschiedenen Korpora angewendet wurden. Der Koeffizient stellt die Stabilität der Verfahren hinsichtlich der Anwendung auf vergleichbare oder themengleiche Quellen dar.

3037 Wörter und TAZ 2123 Wörter – fällt auf, dass die Daten der TAZ Dokumente enthalten, die im Mittel ca. 30% kürzer sind.

Bei den globalen Topic-Modellen ist die Abhängigkeit der Themenanzahl von einem Korpus nur durch das HDP-LDA Verfahren feststellbar, da die LDA Berechnungen von einer festen Vergabe der Themenanzahl ausgehen. Die Abweichungen der Themenanzahl sind bei der Anwendung der HDP-LDA höher, als bei der Verkettung der Topic-Modelle. Die Dokumentlänge scheint auch hier eine Rolle zu spielen, da bei den $HDP_{\widehat{G}}$ Modellen weniger Themen im TAZ-Korpus gefunden werden. Die Lemmatisierung bzw. Grundformreduktion scheint bei den globalen HDP-LDA Modellen keine große Rolle bei der Themenanzahl zu spielen.

Zum einen ist die Anzahl der relevanten Themen für die Thematisierungen LIB und JAP für die Verfahren innerhalb eines Korpus verschieden. Zum anderen ist die die Anzahl der relevanten Themen innerhalb eines Verfahrens für unterschiedliche Korpora nicht einheitlich. Dies lässt den Schluss zu, dass die Abstraktion der Themen durch die Vielfältigkeit der damit verbundenen Informationen in einem Korpus bestimmt wird. Die Art und Weise in der über die Themen berichtet wird, unterscheidet sich, besonders bei Nachrichtenmedien, innerhalb verschiedener Textquellen. Beispielweise finden sich im Guardian mehr relevante Themengruppen für das Thema LIB als für JAP. Das Verhältnis in SZ und TAZ ist hingegen genau umgekehrt. Somit kann angenommen werden, dass die vielfältige oder heterogene Behandlung eines Themas mehr relevante Themengruppen hervorruft. Dies lässt einen Schluss darüber zu, welche Themen intensiver oder abwechslungsreicher aufgenommen werden und durch mehr sekundäre Thematisierungen ergänzt werden.

Zur Robustheit und Reliabilität ist bei den besprochenen Verfahren zusammenfassend festzustellen, dass unterschiedliche Ergebnisse bei der Anwendung gleicher Verfahren in unterschiedlichen aber thematisch vergleichbaren Textquellen auf

- verschiedene Dokumentlängen,
- unterschiedliche morphologische Vorverarbeitung und
- eine verschiedene Nutzung und Akzentuierung der Themen in verschiedenen Textquellen

zurückzuführen sind. Wird zur Messung einer Thematisierung das Material unterschiedlicher Quellen verwendet, müssen vor der Auswertung alle genannten Randfaktoren überprüft werden, um unterschiedliche Ergebnisse einordnen zu können. Grundsätzliche Annahmen können durch vereinzeltes Lesen themenspezifischer Dokumente plausibilisiert werden. Können grundsätzliche Unterschiede in Quantität und Inhalt in den festgestellten relevanten Themen unterschiedlicher Quellen erklärt werden, kann die Reliabilität durch die Analyse der Berichterstattungsmenge in den verschiedenen Korpora zusätzlich abgesichert werden. Dafür wird aus den relevanten Themen, vorzugweise die Hauptthemen, eines jeden Korpus die Dokumenthäufigkeit ermittelt und in einen Längsschnitt abgetragen. Der übereinstimmende Verlauf dieser Längsschnitte des gleichen Themas aus unterschiedlichen Korpora stellt eine weitere Absicherung dar, dass mit einem gewählten Verfahren ähnliche Zusammenhänge in verschiedenen Textquellen gemessen werden. Dieser Vergleich wird in Abschnitt 4.3 für die Thematisierungen LIB und JAP anhand der verschiedenen Verfahren und Korpora durchgeführt, um die Reliabilität an einem konkreten Beispiel zu bewerten.

4.3 Themenhäufigkeiten

In diesem Abschnitt werden die identifizierten relevanten Themen herangezogen, um deren jeweilige Bestandteile und Anteile innerhalb der betrachteten Korpora zu bestimmen. Eine solche detaillierte Analyse im engen Fokus eines Themas findet in der zweiten Phase einer explorativen Analyse statt. Einerseits werden für alle Korpora Anteile ohne Beachtung der Zeitstempel bestimmt. Dadurch wird der Anteil an der Gesamtberichterstattung sichtbar gemacht. Andererseits werden, unter Beachtung der Zeitstempel, diachrone Analysen der Themenanteile durchgeführt. Diese Messungen dienen der Evaluierung und dem weiteren Nachweis der Validität und Reliabilität der Verfahren. Die Evaluierung stützt sich auf die Annahme, dass die

Berichterstattung für Ereignisse und Themen, die eine hohe Relevanz und Einschlägigkeit besitzen, ähnlichen Verläufen innerhalb unterschiedlicher Korpora folgen. Dies zeigt die Forschung der Konsenzanalyse, deren Inhalt die Messung von Themenübereinstimmungen unterschiedlicher Medien ist. Nach einer exemplarischen Studie ist der Anteil der Übereinstimmungen bei Themen mit hoher Aktualität und überregionaler Relevanz erstaunlich hoch (Top, 2006; Merten, 2001).

4.3.1 Häufigkeiten ohne Beachtung der Zeitstempel

Die Messung der Gesamtbestandteile kann für jedes Verfahren unabhängig durchgeführt werden. Der Anteil wird durch zwei unterschiedliche Kennzahlen bestimmt. Einerseits wird der Anteil der Dokumente, die einem Thema zugeordnet sind, durch Formel 3.9 aus Abschnitt 3.4 bestimmt. Andererseits wird der Anteil der Wortformen, die einem Thema zugeordnet sind, mit Formel 3.10 bestimmt. Beide Berechnungsarten können über die Summe aller Dokumente im Korpus bzw. über die Summe aller Token, siehe Formel 3.11 und 3.12, normalisiert werden. Im Fall der Topic-Modelle können einem Dokument mehrere Themen zugeordnet werden. Um zu bestimmen, ob ein Dokument einen relevanten Anteil eines Themas enthält, wird ein Schwellwert festgelegt. Für die folgenden Messungen wird dieser Schwellwert auf 25% festgelegt. Welche Anteile die Themen an einem Dokument haben, wird für jedes Dokument aus der Verteilung $p(\mathbf{z}|d)$ abgelesen.

In Tabelle C.1, die im Anhang dargestellt ist, wurden die relative Themenanteile für alle berechneten Verfahren in allen Korpora mit Hilfe der Formeln 3.11 und 3.12 bestimmt. Eine erwartbare Beobachtung ist, dass die Themenanteile unter den verschiedenen Korpora variieren. Die Variation der Themenanteile ist auch in unterschiedlichen Verfahren oder Parametrisierungen zu beobachten, die auf einen gleichen Korpus angewendet werden. Diese Beobachtung betrifft insbesondere die Topic-Modelle, durch deren Parametrisierung das Analyseergebnis signifikant beeinflusst werden kann. Die Themenanteile für die Thematisierung JAP_all, berechnet aus den Verfahren $\text{LDA}_{\widehat{G}}$, $\text{LDA}_{\check{G}}$, $\text{LDA}_{0.5}$ und $\text{LDA}_{0.6}$, sind beispielsweise bei 0.087, 0.049, 0.075 und 0.067. Ausgedrückt in absoluten Dokumentzählungen entspricht dies 719, 408, 620 und 533. Es wird deutlich, dass die Abstraktion bzw. die Granularität der Themen, die durch die einzelnen Verfahren erzeugt werden Auswirkung auf die Menge der zugeordneten Wörter und Dokumente haben. Je grober die Thematisierung vorgenommen wird, desto größer ist die Auswahl an Dokumenten, an denen ein Thema einen Anteil hat. Dies wird beim Vergleich der Verfahren $\text{LDA}_{\widehat{G}}$ und $\text{LDA}_{\check{G}}$ deutlich. Die Trennung in viele fokussierte Themen in $\text{LDA}_{\widehat{G}}$ führt dazu, dass ver-

wandte Aspekte, wie andere mit Erdbeben verbundene Ereignisse, nicht mit beachtet werden. Die Dokumentzählung ist geringer als bei einer abstrakteren Betrachtung der Themen. Diese Beobachtung erscheint logisch, da in jedem Ressort, was eine sehr abstrakte Themendefinition in Tageszeitungen darstellt, über unterschiedliche damit verbundene und spezifische Ereignisse berichtet wird. Dies stellt die Reliabilität oder Validität der Topic-Modellen nicht infrage. Vielmehr ist jede Parametrisierung innerhalb der Anwendung von Topic-Modellen als Einzelverfahren zu sehen, welches andere Ergebnisse für andere Analysezwecke erzeugt. Die Reliabilität und Validität hängt stark von der Fragestellung ab, auf deren Bedürfnisse Topic-Modelle durch die Parametrisierung flexibel regieren können.

Die Betrachtung der Gesamtanteile in einem Korpus zeigt nicht, wie sich die Dokumente über die Zeitscheiben verteilen. Bei zwei Zählungen von 719 und 408, wie oben beschrieben, ergibt sich die Differenz von 311 Dokumenten. Bei einer Betrachtung von 61 Tagen macht das lediglich einen Unterschied von durchschnittlich 5 Dokumenten am Tag aus. Die Unterschiede lassen sich mit dieser Betrachtungsweise besser einordnen. Es ist nachvollziehbar, dass die Zuordnung der Dokumente um diesen Bereich abweichen kann, wenn sich der Abstraktionsgrad ändert. Für die Anwendung der TDT kann mit Hilfe der Themenanteile keine Aussage über die Validität gemacht werden. Die Themenanteile repräsentieren lediglich die Verbreitung eines Verwendungszusammenhangs mit einer bestimmten Abstraktion. Dies zeigt, dass die Beurteilung der Validität der vorgeschlagenen Verfahren nicht anhand der reinen Themenanteile, ohne Beachtung der Zeitstempel, beurteilt werden kann. Die Verfahren müssen vielmehr die richtige Themenauflösung für eine gegebene Fragestellung bereitstellen, was zu unterschiedlichen Häufigkeiten und Anteilen eines Themas führen kann. Die Beurteilung der Reliabilität und Validität muss demnach mit anderen Methoden erfolgen.

4.3.2 Häufigkeiten mit Beachtung der Zeitstempel und Evaluation

Zur Beurteilung der praktischen Leistungsfähigkeit von Verfahren dient in der Informatik meist eine empirische Evaluation. Mit den Ergebnissen kann beurteilt werden, in welcher Qualität eine Aufgabe hinsichtlich einer Testmenge oder eines Gold-Standards erfüllt wird. Gold-Standards sind in der Regel manuell beurteilte Daten, wobei Anwender das gewünschte Ergebnis eines Verfahrens, wie beispielsweise eine Klassifikation, manuell erstellen. Für unüberwachte Verfahren zur Identifikation von Themen in diachronen Textquellen existieren Gold-Standards, die für die Beurteilung herangezogen werden können (Allan, 2002, vgl. S. 17). Wenn die Erstellung von Themenlisten aus Textkollektionen als inhaltsanalytische Aufgabe im kommunikationswissenschaft-

lichen Sinn betrachtet wird, ist die Ergebnisproduktion allerdings immer abhängig von der Definition des Forschungsprozesses. Dies liegt daran, dass die Themen durch Fachanwender definiert werden und Themenzusammenhänge unterschiedliche Abstraktionsgrade haben können. Für die Evaluation der Gültigkeit der Verfahren, im Hinblick auf die Anwendung in der Inhaltsanalyse, ist ein fester Gold-Standard, in Form einer statischen Testmenge, ungeeignet. Die Verfahren, die in Kapitel 3 besprochen werden, sind durch die Wahl der Parameter in der Lage unterschiedliche Abstraktionsgrade abzubilden. Die Erstellung einer Testmenge mit zugehörigen Themen und Dokumenten ist damit immer abhängig von der Parameterwahl innerhalb der Verfahren und den Analysezielen. Aus diesen Gründen müssen andere Mittel für die Evaluation gefunden werden.

Ein wichtiger Bestandteil kommunikationswissenschaftlicher Themenanalysen ist die Bestimmung der Nachrichtenfaktoren einer Nachricht.[9] Je mehr Nachrichtenfaktoren innerhalb einer Meldung erfüllt werden, desto höher ist der Nachrichtenwert und die Meldung enthält entsprechend mehr Aufmerksamkeit in der Öffentlichkeit. Eine Thematisierung innerhalb einer diachronen Textquelle, vor allem aber in Nachrichtenquellen, beinhaltet Dokumente mit unterschiedlichem Nachrichtenwert. Dieser orientiert sich vor allem an Ereigniszusammenhängen oder Personenbezügen. Zusätzlich ist bei einer andauernden Thematisierung eine Nachricht über das Thema selbst mit hoher Aufmerksamkeit verbunden. Wird der Themengehalt eines Korpus mit automatischen Methoden bestimmt, um damit eine Themenanalyse im kommunikationswissenschaftlichen Sinn zu erstellen, so müssen Größen wie Nachrichtenfaktoren und die öffentliche Aufmerksamkeit möglichst gut interpretiert werden können. Vor allem Ereignisse oder Personenbezüge sorgen für hohe Nachrichtenwerte, da sie gleichzeitig Überraschung, Konflikt- oder Schadpotential und eine Personalisierung mit sich bringen. Somit muss die Menge der Dokumente oder Wörter, die einem Thema in einem bestimmten Zeitraum zugeordnet sind, mit dem Auftreten von Ereignissen und dem öffentlichen Agieren von Organisationen und Personen, die mit diesem Thema in Zusammenhang stehen, korrelieren. Die Aufmerksamkeit und die Menge der Berichterstattung, die einem Thema gewidmet wird, muss demnach mit diesen Nachrichtenfaktoren korrelieren. Eine Evaluierung der Validität einer automatischen Themenanalyse kann durchgeführt werden, indem die Zeitreihen der Häufigkeiten, die aus den automatisch erstellten Themen generiert werden können, mit realen Ereignissen verglichen werden. Wenn die generierten Häufigkeiten (Dokumente, Wörter)

[9] In Abschnitt 2.1.3 auf S. 32 werden die Faktoren vorgestellt.

einer Thematisierung mit Schlüsselereignissen genau dieser korrelieren, so sind valide Schlüsse und Analysen möglich. Dieses Verhalten wird umfassend in der hier dargestellten Evaluierung getestet. Bestätigt die Evaluierung, dass dieser Zusammenhang zwischen realen Ereignissen und automatisch gefundenen Themen hergestellt werden kann, so sind die Zeitreihen und Beurteilungen, die aus den automatisierten Analysen resultieren auch für andere Themen valide.

Für die Evaluierung werden die prominenten Themen „Fuskushima“ (JAP) und „Lybien“ (LIB) gewählt, da aus entsprechenden Jahreschroniken Ereignisse, Personen und Organisationen, die eine Rolle für diese Themen spielen, manuell extrahiert werden können.[10] Die Auswahl der Referenzthemen fiel auf sehr prominente Themen, um die Vollständigkeit der zusammengetragenen Ereignisse zu gewährleisten. Die Quellenlage ist sehr gut und die Auswahl der Ereignisse ist dadurch intersubjektiv nachvollziehbar.

Reliabilität

Um die Reliabilität der Verfahren zu beurteilen, muss bestimmt werden, wie konstant die Messung bei mehrmaliger Anwendung eines Verfahrens auf einen Korpus ist. Für die Messung mit TDT erübrigt sich ein Test, da das Verfahren ein Clustering der Dokumente nach einer festen Regel vornimmt. Bei mehrmaliger Anwendung auf einen Korpus ist die Zuordnung der Dokumente übereinstimmend. Bei der Anwendung der Topic-Modelle ist dies allerdings nicht der Fall. Die Inferenz der Modelle basiert auf Variational Bayes Methoden oder Gibbs Samplern. Beide Varianten nähren sich den optimalen Modellparametern nur an. Die Verfahren unterliegen Zufallsprozessen, wobei ein wesentlicher Bestandteil aus der Verwendung von Wahrscheinlichkeitsverteilungen besteht. Die Lösungsverfahren sind von Zufallsprozessen abhängig. Diese damit geschätzten Optima können jedoch lokal sein, sodass mehrere optimale Lösungen existieren (Griffiths u. Steyvers, 2004). Aus diesem Grund muss bestimmt werden, wie stabil die mehrmalige Messung mit Topic-Modellen unter konstanten Vorbedindingungen an einem Korpus ist. In einigen Arbeiten wird die Stabilität von Topic-Modellen stark infrage gestellt. In Koltsov u. a. (2014) wird ein Experiment durchgeführt, welches mehrere gleichartig erstellte Modelle miteinander vergleicht. Die Autoren beziehen sich auf einen Themenvergleich, der auf einer normalisierten Variante der Kullback-Leibler Divergenz (KL) basiert. Die Unterschiede der Themen, die aus mehreren Inferenzen resultieren, können so quantifiziert werden. Die Autoren

[10] Die Schlüsselereignisse und Ereignisbezüge wurden aus den folgenden Quellen erstellt: (Handelsblatt; Zeit; Tagesspiegel 1; Tagesspiegel 2; Spiegel; Wiki 2; Wiki 1; Wiki 3).

stellen fest, dass nur bei einem Wert der KL zweier Themen von > 0.93 sicher ist, dass die Themen ähnliche Zusammenhänge repräsentieren. Bei einer KL von 0.85 zwischen 2 Themen stellen sie fest, dass die Sortierung und Wahrscheinlichkeiten der Themen schon erheblich voneinander abweichen. Deshalb kommen sie zu dem Schluss, dass weniger als die Hälfte aller Themen, im Hinblick auf die KL-Werte, als ähnlich anzusehen sind und die produzierten Themen aus mehreren gleichartig berechneten Modellen nicht stabil berechnet werden. In Niekler u. Jähnichen (2012) wird hingegen gezeigt, dass die Anwendung von Abstandsmaßen auf den vollständigen Verteilungen $p(\mathbf{w}|z)$ ähnliche und nicht verwandte Themen schlecht voneinander diskriminiert. Dies liegt daran, dass ein Teil der Verteilungen mit wenig Wahrscheinlichkeitsmasse, der sogenannte Long-Tail, für alle Themen ähnlich strukturiert ist. Der Long-Tail kann deshalb als eine Art Rauschen der Zufallsprozesse in den Inferenzverfahren angesehen werden. Durch die hohe Übereinstimmung der Struktur des Long-Tail wird eine Ähnlichkeit suggeriert, wenn die Abstandmaße direkt angewendet werden, obwohl sich die Themen inhaltlich unterscheiden. Diesen Effekt spricht Koltsov u. a. (2014) an und es ist schwer, Themenzugehörigkeiten sicher zu unterscheiden. Aus diesen Gründen arbeitet der Test, ob ein Thema ähnlich zu einem anderen Thema ist, unter diesen Bedingungen nicht zuverlässig.

Die in Abschnitt 3.2.4 vorgeschlagene Vorgehensweise, dass nur wenige hochwahrscheinliche Bestandteile einer Verteilung für den Vergleich genutzt werden, führt zu einer besseren Unterscheidbarkeit von ähnlichen und nicht ähnlichen Themen. Aus diesem Grund wird die Reliabilität an dieser Stelle durch die Vergleichsstrategie aus Niekler u. Jähnichen (2012) noch einmal überprüft. Das Experiment berechnet 10 Topic-Modelle (LDA) mit 50 Themen für einen kleinen Korpus unter exakt gleichen Bedingungen. Als zu testende Dokumentmenge wird jeweils der 15.03.2011 aus den Korpora TAZ und Guardian gewählt. Die 10 Modelle für jedes Korpus werden jeweils untereinander verglichen. Dafür werden, wie bei der Verkettung der Themen, nur die hochwahrscheinlichen 10 Terme eines Themas genutzt. Als Ähnlichkeitsmaß wird die Kosinusdistanz genutzt.

Für jedes Thema eines Modells wird nach Prozedur 3, die im Anhang dargestellt ist, ermittelt, welches Thema des zu vergleichenden Modells die höchste Ähnlichkeit aufweist. Diese maximale Ähnlichkeit eines Themas in einem Modell zu einem Thema in einem anderen Modell wird in einer Liste vermerkt. Dies wird für alle Themen in einem Modell durchgeführt, sodass für jedes Modell eine Liste von 50 Ähnlichkeiten ermittelt wird. Anhand der maximalen Ähnlichkeiten der Themen eines Modells zu einem anderen Modell kann nun ermittelt werden, wie viele Themen als ähnlich an-

zusehen sind. In der vorangegangenen Untersuchung wird festgestellt, dass bei einer Ähnlichkeit von 0.4 Themen verbunden werden, die teilweise nicht zusammengehören oder andere Aspekte eines Themas betreffen. Dieser Effekt kann bei einer Mindestähnlichkeit von 0.5 bei den Themen JAP und LIB nicht mehr feststellt werden. Aus diesem Grund werden in der Liste der maximalen Ähnlichkeiten die Elemente ermittelt, die den Schwellwert von 0.5 nicht erreichen. Es werden die Themen gezählt, die im direkten Vergleich mit einem anderen Modell nicht zugeordnet werden können. Dies wird für alle Kombinationen der 10 berechneten Modelle wiederholt, sodass am Ende eine Matrix entsteht, in welcher die Anzahl der nicht zuordenbaren Themen für jeden Vergleich zweier Modelle eingetragen werden. Aus der oberen und der unteren Dreiecksmatrix lässt sich nun ein Mittelwert errechnen. Dieser Mittelwert repräsentiert die durchschnittlich nicht stabilen Themen, die bei der Inferenz einer LDA entstehen. Für den Testkorpus aus der TAZ ergibt sich ein Fehler von 18%, wenn die Anzahl der nicht zuordenbaren Themen normiert wird. Somit können nach dieser Evaluierung, 10 gleich berechneter Modelle, 82% aller Themen als stabil angesehen werden. Das Korpus, welches aus der Online-Ausgabe des Guardian erstellt wurde, weist einen Fehler von 23% bei 10-facher Berechnung eines LDA Modells auf und liegt damit bei einer Stabilität von 77%. Im Sinne der Inhaltsanalyse sind diese Werte für die Reproduzierbarkeit einer Verfahrensanwendung als akzeptabel und brauchbar anzusehen. Die Inferenz der HDP basiert auf ähnlichen Mechanismen, wie die der LDA, und die Ergebnisse eines solchen Reliabilitätstests sind ähnlich zu denen der LDA.

Validität

Die Reliabilität eines Verfahrens zeigt nicht, ob die erzeugten Ergebnisse valide sind. Da die erzeugten Themen die Beantwortung einer Fragestellung sicher stellen müssen, ist die Bewertung der Validität anhand einer konkreten Fragestellung vorzunehmen. Je nach thematischer Breite müssen die Ergebnisse einem gewünschten Abstraktionsgrad entsprechen. Grundsätzlich sollten die Themen aber immer in Abhängigkeit zu externen Ereignissen stehen, die eine Berichterstattung über ein Thema auslösen. Nachfolgend soll der Nachweis von Validität der verwendeten Verfahren geführt werden. Es werden zwei Ideen zur Evaluierung diskutiert. In Abschnitt 2.1.3 werden Nachrichtenfaktoren und die zeitliche Darstellung von Themenintensitäten besprochen. Es wird gezeigt, dass sich die Berichterstattungsmenge, und demnach das angenommene öffentliche Interesse, an Nachrichtenfaktoren und themenzugehörigen Schlüsselereignissen orientiert. Zeitreihen der Berichterstattungsmenge eines Themas

zeigen Zeitpunkte für die Analyse von Nachrichtenfaktoren oder für die Identifikation von Schlüsselereignissen auf, da eine erhöhte Berichterstattung mit diesen Faktoren korrelieren soll. Auf der anderen Seite kann durch die externe Bestimmung von Nachrichtenfaktoren oder Schlüsselereignissen evaluiert werden, ob eine in den Textdaten automatisch gemessene Themenintensität hinsichtlich externer Ereignisse oder nachweislich vorhandener Nachrichtenfaktoren valide abgebildet ist. Die Zeitreihen helfen, Ereignisse und wichtige Faktoren einer Thematisierung zu finden oder anhand der Ereignisse zu testen, ob die Häufigkeiten in den Zeitreihen korrespondieren und den richtigen Zusammenhang abbilden. Die Evaluierung der Validität von Themenmessungen in digitalen Textkorpora kann demnach durch Nachrichtenfaktoren oder externe Ereignisse erfolgen. Für die Evaluierung mit Nachrichtenfaktoren muss festgelegt werden, welchen Nachrichtenwert vorhandene Nachrichtenfaktoren induzieren, sodass die Ist-Nachrichtenintensität mit einer Soll-Nachrichtenintensität verglichen werden kann. Einer Untersuchung zufolge kann gezeigt werden, dass der Wert einer Nachricht aber nicht immer mit dem Vorhandensein oder der Kombination von Nachrichtenfaktoren einhergeht (Kepplinger, 2011). Die eigentliche Selektionsentscheidung für eine Nachricht basiert demnach nicht auf Nachrichtenfaktoren, wohl aber deren Platzierung und Umfang. Demnach kann die Analyse der Nachrichtenfaktoren mit einer Analyse der Platzierung der Inhalte verknüpft werden. Dies erlaubt allerdings keine Rückkopplung auf die Selektionsentscheidungen der Journalisten, die einzelne Beiträge erstellt haben. Somit können Nachrichtenfaktoren zwar genutzt werden, um die Relevanz bereits selektierter Nachrichten zu analysieren, für die Klärung der Validität einer Themenanalyse selbst kann die Analyse der Nachrichtenfaktoren aber nicht dienen. Bei Schlüsselereignissen ist der Zusammenhang zur Nachrichtenselektion stärker, da Schlüsselereignisse neue Themen schaffen oder ein Thema grundsätzlich neu strukturieren (Rauchenzauner, 2008, vgl. S. 24). Damit einher geht das Bedürfnis, sich erneut über ein Thema zu informieren. Daraus resultiert eine höhere Aufmerksamkeit, die einem Thema zukommt. Die Außergewöhnlichkeit eines Schlüsselereignisses spielt eine große Rolle bei der Selektionsentscheidung von Journalisten. Die Bedeutung eines Schlüsselereignisses bezieht sich auch auf den Beitrag neuer Sachverhalte oder die Tragweite des Ereignisses. Für den hier angestrebten Nachweis der Validität der Themenmessungen lässt sich diese Eigenschaft der Schlüsselereignisse nutzen, um die extrahierten Zeitreihen aus LDA, HDP und TDT zu evaluieren. Aus den externen Quellen oder Chroniken müssen Schlüsselereignisse für die Thematisierungen JAP und LIB identifiziert werden. Bilden die aus den automatischen Verfahren extrahierten und identifizierten Themen diese Schlüsselereignisse ab, so kann davon

Thematisierung JAP
11. März – Erdbeben (Tōhoku-Erdbeben)
12. März – Explosion Block I
14. März – Explosion Block III
15. März – Explosion Block II und IV
16. März – Vermutung einer Kernschmelze
28. März – Kernschmelze wird zugegeben
7. April – Die Region wird von einem Erdstoß (2011 Miyagi-Präfektur Erdbeben) der Stärke 7,1 erschüttert.
11. April – erneuter Erdstoß (Fukushima Hamadōri Erdbeben) 7,1
Thematisierung LIB
3. März – Rebellen fordern UN-Flugverbotszone
7. März – Übergangsrat drängt auf internationale Anerkennung
10. März – Übergangsrat durch Frankreich anerkannt
17. März – Einrichtung der internationalen Flugverbotszone (UN-Resolution 1973)
19. März – US-Raketen-Angriffe
20. März – Politische Reaktionen auf die Militäraktionen
21. März – internationale Kritik an UN-Einsatz

Tabelle 4.5: Auswahl der Schlüsselereignisse zu den Themen JAP und LIB im Zeitraum März - April 2011.

ausgegangen werden, dass die Verfahren korrekte Sachverhalte messen. Die Aufnahme von Schlüsselereignissen ist nicht abhängig von mehreren Nachrichtenfaktoren, sondern tatsächlich nur von der besonderen Ausprägung eines oder weniger Faktoren. Zusätzlich gilt für Schlüsselereignisse, dass die Selektionsfunktion der Journalisten und Redaktionen aufgehoben wird und die Ereignisse unabhängig von dem Interesse, der Ideologie oder thematischen Ausrichtung einer Redaktion übernommen werden und die Gatekeeper-Funktion des Jounalisten lediglich über die Ausrichtung eines Beitrages entscheidet (Rauchenzauner, 2008, vgl. S. 179). Demnach sollte durch das erhöhte Informationsinteresse der Öffentlichkeit und die erhöhte Aufnahmebereitschaft unterschiedlicher Medien sichergestellt sein, dass Schlüsselereignisse immer zu einer erhöhten Berichterstattung führen.

Die Idee für die Evaluierung liegt nun darin, externe Schlüsselereignisse zu identifizieren und mit den gemessenen Themenverläufen aus den automatischen Verfahren zu korrelieren. Die Schlüsselereignisse, welche für die externe Validierung verwendet werden, sind in Tabelle 4.5 aufgelistet.[11]

Die Ereignisse, die eine Thematisierung begleiten, sind nicht alle als Schlüsselereignis zu betrachten. Vor allem die Analyse von Themen, die nicht im lokalen Kontext zum Erscheinungsort eines Mediums stehen, müssen zwischen Ereignissen lokaler und internationaler Bedeutung unterscheiden. Im Fall der hier gezeigten Analyse werden Nachrichtenmedien analysiert, die keinen direkten lokalen Zusammenhang zu den Themen und Ereignissen haben. Vielmehr werden die Ereignisse in einem lokalen

[11] Die Ereignisse wurde mit Hilfe der Quellen (Handelsblatt; Zeit; Tagesspiegel 1; Tagesspiegel 2; Spiegel; Wiki 2; Wiki 1; Wiki 3) definiert.

Kontext der Nachrichtenmedien von der Öffentlichkeit diskutiert. Beispielsweise wird der Umgang mit Kernenergie nach der Katastrophe in Japan in lokalen Kontexten in Europa diskutiert. Dennoch existieren für die meisten Themen Schlüsselereignisse, die eine internationale Relevanz haben. Um für die Untersuchung geeignete Schlüsselereignisse zu wählen, wird die Definition aus Rauchenzauner (2008, vgl. S. 17) herangezogen. Demnach muss ein Schlüsselereignis das Informationsbedürfnis erhöhen, spektakulär sein, einen bestehenden Diskurs oder ein Thema grundsätzlich verändern und eine große Tragweite haben. Die bloße Zustandsbeschreibung einer Thematisierung kann nicht als Schlüsselereignis gelten. Die Ereignisse für die Thematisierungen wurden aus mehreren Chroniken zusammengetragen. Für die Thematisierung JAP bestehen die Meldungen aus Berichten über neue Beben, Strahlungswerte, Zustände der Reaktoren oder Pressemitteilungen beteiligter Akteure. Nach der Definition für Schlüsselereignisse kann das Erdbeben selbst als spektakuläres Ereignis gelten, welches einen hohen Schaden anrichtet. Die Explosion der Reaktoren verändert die Thematisierung selbst, da das Erdbeben und seine Folgen ab diesem Zeitpunkt als Nuklearkatastrophe beschrieben werden. Da teilweise nicht klar ist, was in den Reaktoren passiert, modifizieren Meldungen über eine mögliche Kernschmelze das Thema, da die Katastrophe dadurch viele unbekannte Faktoren erhält, über die sich die Öffentlichkeit informieren will. Meldungen über neue Erdstöße sind als Schlüsselereignis zu sehen, da diese ähnliche Schäden anrichten können und immer Bedeutung für den weiteren Verlauf des Themas haben. Die anhaltenden Berichte über Schäden, Strahlung, Aufräumarbeiten und Hilfen können als Statusberichte gesehen werden, die den Inhalt der Thematisierung nicht wirklich verändern. Im Fall der Thematisierung LIB kann die Transformation eines lokalen zu einem internationalen Konflikt beobachtet werden. Berichte über Erfolge der Rebellen oder der Regierungstruppen begleiten den Konflikt als Statusmeldungen. Eine wirklich neue Dimension bekommt die Thematisierung durch die Forderung der Rebellen nach internationaler Hilfe und die Forderung nach einer Flugverbotszone. Dadurch muss dieser lokale Konflikt durch die internationale Gemeinschaft diskutiert werden. Gleichzeitig kämpft der Übergangsrat in Libyen um internationale Anerkennung. Die Einrichtung einer Flugverbotszone stellt eine politische Entscheidung dar, die dem Konflikt endgültig eine Wende gibt. Die lokalen Auseinandersetzungen werden danach durch erste Angriffe durch US-Streitkräfte zu einem internationalen Konflikt. Diese Ereignisse ziehen die internationale Bewertung der Militäraktionen mit sich, was das Informationsinteresse der Öffentlichkeit erhöht. Auch hier werden deshalb die Schlüsselereignisse von Ereignissen getrennt, die Sta-

tusmeldungen beinhalten, aber nicht die Qualität der Thematisierung grundsätzlich verändern.

Um die semi-automatisch erstellen Themen und deren Dynamik, dargestellt als Zeitreihe der Publikationshäufigkeit, nun mit den Schlüsselereignissen zu vergleichen, werden diese in eine Zeitreihe überführt. Dafür wird jedes Schlüsselereignis zunächst mit dem Wert 1 für die jeweilige Zeitscheibe in einer Zeitreihe dargestellt. Die Schlüsselereignisse für eine Thematisierung werden formuliert über eine Menge T, die aus Zeitstempeln t zusammengesetzt ist. Somit wird für alle festgelegten Schlüsselereignisse $t \in T$ in einer künstlichen Zeitreihe $SE_t = 1$. In einer solchen künstlichen Zeitreihe sind Toleranzen für die Zeitpunkte der Schlüsselereignisse erlaubt, da die Verbreitung der Information über ein Ereignis länger anhalten kann. Die Intensität nimmt aber steig ab. Es wird für jeden Zeitpunkt $t \in T$ definiert, dass die folgenden Zeitscheiben als fallende Funktion dargestellt werden. Für den Zeitpunkt eines Schlüsselereignisses wird die Zeitreihe auf $SE_{t+1} = 0.5$, $SE_{t+2} = 0.25$ und $SE_{t+3} = 0.1$ für alle $t \in T$ gesetzt. Falls innerhalb dieses modellierten Abklingens ein weiteres Schlüsselereignis stattfindet, so wird das Abklingen eines vorherigen Schlüsselereignisses an dieser Stelle mit dem Einfluss des neueren Ereignisses überschrieben. Die so definierten künstlichen Zeitreihen für beide untersuchten Thematisierungen werden in den Abbildungen 4.10 und 4.9 als Längsschnitt dargestellt.

$$SE_t = \begin{cases} 1 & t \in T \\ 0.5 & t \notin T \wedge (t-1) \in T \\ 0.25 & t \notin T \wedge (t-1) \notin T \wedge (t-2) \in T \\ 0.1 & t \notin T \wedge (t-1) \notin T \wedge (t-2) \notin T \wedge (t-3) \in T \\ 0 & \text{sonst} \end{cases} , \tag{4.1}$$

In der empirischen Medienwissenschaft gehören Längsschnittanalysen von Termhäufigkeiten oder Themenintensitäten zum Methodenrepertoire. Dabei spielt die Analyse von externen Abhängigkeiten der Längsschnitte eine gesonderte Rolle. Die „Interrupted time series analysis“ nutzt Zeitreihenmodelle wie ARIMA, um signifikante Zusammenhänge zwischen externen Ereignissen und Einflüssen auf Intensitätsverläufe zu bestimmen, wobei externe Ereignisse ebenfalls als Zeitreihe dargestellt werden (McDowall, 1992). Zum Ereigniszeitpunkt werden die Werte in den Zeitreihen auf den 1 gesetzt. Der Rest wird mit dem Wert 0 festgelegt. Die Länge der Zeitreihe muss der Länge der analysierten Zeitreihe entsprechen. Das Verfahren bestimmt, ob es zwischen dem Ereignis und der Zeitreihe aus einem Thema oder einer Worthäufigkeit einen Zusammenhang gibt. Dieses Vorgehen kann allerdings nur jeweils

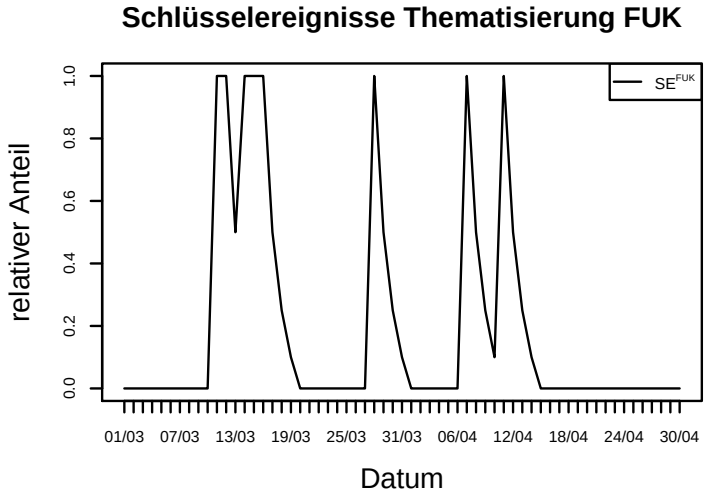

Abbildung 4.9: Künstliche Zeitreihe der Thematisierung JAP mit Abklingeigenschaften der Schlüsselereignisse.

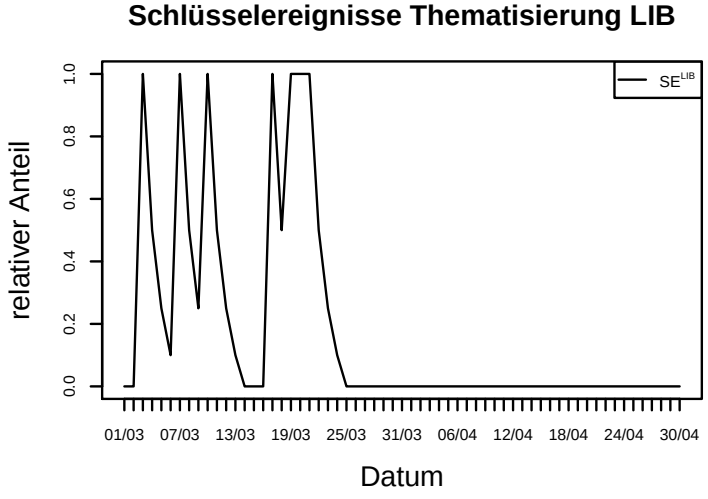

Abbildung 4.10: Künstliche Zeitreihe der Thematisierung LIB mit Abklingeigenschaften der Schlüsselereignisse.

1 Ereignis testen. Jeder Zusammenhang zu einem Ereignis muss separat überprüft werden. Da die künstlichen Zeitreihen die gleiche Länge haben wie die extrahierten Längsschnitte der Themenintensitäten, bietet sich die Pearson-Korrelation bzw. die Stichprobenkorrelation für den Vergleich an. Diese misst den linearen Zusammenhang bzw. die lineare Abhängigkeit zweier Zufallsvariablen. Die Evaluierung der Zeitreihen besteht aus einem statistischen Test, der die Unabhängigkeit der künstlichen Zeitreihen und der Themenzeitreihen testet. Die Messung kann als valide angesehen werden, wenn diese angenommene Hypothese der Unabhängigkeit verworfen werden kann. Die Selektion und die Veröffentlichung von Nachrichtenartikeln in Tageszeitungen unterliegen allerdings einer Verzögerung, die durch die Produktionszeit oder die Erscheinungstermine der Organe bedingt sind. Somit würde es wenig Sinn machen, die Ereignisse direkt mit den Zeitreihen aus den analysierten Korpora zu vergleichen. Aus diesem Grund wird der Test der Zeitreihen über eine Kreuzkorrelation durchgeführt. Die Datenpunkte der künstlichen Zeitreihen werden auf der Zeitachse schrittweise verschoben und die Korrelation der verschobenen Zeitreihe mit einem Original wird bestimmt. Die Verschiebung wird "lag„ genannt und die Bildung der Korrelation über alle möglichen Verschiebungen wird als Kreuzkorrelationsfolge bezeichnet (Schüßler, 2008). Daraus ergibt sich eine Funktion anhand derer abgelesen werden kann, ob bei einer bestimmten Verschiebung der Zeitreihen zueinander eine maximierte Korrelation existiert. Für den hier durchgeführten Test wird die schrittweise Verschiebung auf 3 begrenzt, da davon ausgegangen werden kann, dass über die Schlüsselereignisse nach maximal 3 Tagen in den analysierten Korpora berichtet wird. Aus der Messung der Korrelation aus den jeweils 3 Verschiebungen wird die maximale Korrelation selektiert und das "lag„ notiert. Die maximale Korrelation wird einem Test unterzogen, um zu überprüfen, ob die Hypothese gilt, dass beide verglichenen Zeitreihen bzw. Variablen unabhängig sind. Dies wird mit einen t-Test realisiert. Unter der Null-Hypothese, das zwei Variablen, bzw. Zeitreihen unabhängig sind, ergibt sich ein Testwert einer t-verteilten Zufallsvariable mit n-2 Freiheitsgraden $t > t_{n-2;1-\alpha/2}$. Der Vergleichswert ist $t_{61-2;0.975} = 2.00$ bei einem Konfidenzbereich von 95%. Der Test wird folgendermaßen durchgeführt:

- Berechnung der relativen Zeitreihen $P^D_{k,t}$ und $P^W_{k,t}$ für die Thematisierungen JAP und LIB, die aus den Berechnungen der $\text{LDA}_{\widehat{G}}$, $\text{LDA}_{\check{G}}$, $\text{LDA}_{0.5}$, $\text{LDA}_{0.6}$, $\text{HDP}_{\widehat{G}}$, $\text{HDP}_{\check{G}}$, $\text{HDP}_{0.5}$, $\text{HDP}_{0.6}$ und TDT in den Korpora Guardian, SZ, TAZ, SZ_BASE und TAZ_BASE hervorgegangen sind.[12] Dabei sind die Thema-

[12] Für die Zeitreihen werden die Formel 3.15 und 3.16 verwendet.

tisierungen weiterhin in zwei Varianten unterteilt. Eine Thematisierung pro Verfahren und Korpus für die Hauptbestandteile des Themas und eine Thematisierung, welche die Hauptthemen inklusive aller Nebenthemen abbildet. Daraus ergeben sich insgesamt 300 Zeitreihen.

- Erstellung einer Kreuzkorrelationsfolge $r_{v_1,v_2}(\lambda) = \varepsilon\{v_1(k+\lambda)v_2(k)\}, \lambda \in \{0,1,2,3\}$ zwischen den jeweiligen künstlichen Zeitreihen und den 300 semiautomatisch erzeugten Zeitreihen (Schüßler, 2008, vgl. S. 162).

- Selektion des "lag„ in allen Kreuzkorrelationsfolgen, welches eine maximale Korrelation aufweist.

- Berechnung des t-Wertes für alle maximalen Korrelationen $t = \left(r_{v_1,v_2} \times \sqrt{n-2}\right) \Big/ \left(\sqrt{1-r^2_{v_1,v_2}}\right)$ (Hartung, 2009, vgl. S. 548).

- Vergleich der t-Werte mit $t_{61-2;0.975} = 2.00$. Wenn $t > 2.00$ ist, kann die Hypothese der Unabhängigkeit bzw. der Nicht-Korreliertheit für die entsprechende Zeitreihe zurückgewiesen werden.

Die maximale Korrelation der einzelnen Zeitreihen mit der jeweiligen künstlichen Zeitreihe ist im Anhang in Tabelle C.2 verzeichnet. Die Berechnung der einzelnen t-Werte und der Vergleich mit $t_{61-2;0.975}$ zeigt, dass die Unabhängigkeitsannahme bei nur 7 Zeitreihen nicht zurückgewiesen werden kann. Das Konfidenzniveau bei einer unterstellten Unabhängigkeit mit $r_{v_1,v_2} = 0$ beträgt 0.257 bei einem Konfidenz-Niveau von 0.95%. In der Tabelle sind die Korrelationswerte, die diesen Wert nicht überschreiten, abzulesen.

Es ist zu erkennen, dass ausschließlich innerhalb der Thematisierung LIB Zeitreihen erstellt wurden, die nicht abhängig von den Schlüsselereignissen scheinen. Dies betrifft die Zeitreihen $P^D_{k,t}$ für die Verfahren $\text{LDA}_{\hat{G}}$ (SZ/lib_{main}, $\text{GUARDIAN/lib}_{main}$), $\text{HDP}_{\check{G}}$ ($\text{GUARDIAN/lib}_{main}$, $\text{TAZ\_BASE/lib}_{main}$, lib_{all}) und TDT ($\text{TAZ\_BASE/lib}_{main}$, SZ_BASE/lib_{all}). Die Ursache ist in diesen Fällen hauptsächlich in der Definition der Thematisierung zu suchen. Während der Trennung des Themas in Haupt- und Nebenthemen wurde festgelegt, dass die Berichterstattung über den Konflikt in Libyen das Hauptthema ist und die Berichte über die UN-Resolution und die Internationale Gemeinschaft wurden zu Nebenthemen erklärt. Dies hat den einfachen Grund, dass die Haupt- und Nebenthemen nach lokalem und internationalem Bezug auf das Thema getrennt wurden. In einigen Korpora verhält es sich aber so, dass die Trennung der Berichte nach Haupt- und Nebenthemen dazu führt, dass die Schlüsselereignisse, die

innerhalb eines internationalen Kontext stattfinden, nicht in der bloßen Berichterstattung über den Konflikt abgebildet sind. Dadurch sind die schwachen Korrelationen bei den Zeitreihen aus LIB_main zu erklären, da die eigentlichen Schlüsselereignisse in den definierten Nebenthemen abgebildet sind. Es wäre demnach ratsam für die weitere Analyse des Themas LIB mit Haupt- und Nebenthemen zu arbeiten. Weiterhin sind zwei Zeitreihen auf der Basis von LIB_all scheinbar unabhängig von den Schlüsselereignissen. Dies betrifft einmal das Verfahren TDT bei der Anwendung auf den Korpus SZ_BASE und das Verfahren $\text{HDP}_{\tilde{G}}$ bei der Anwendung auf TAZ_BASE. Die Anwendung der Verfahren und die Auflösung der Themen ist immer abhängig von der Wortzusammensetzung und der Struktur der Dokumente in einem Korpus. Für die Trennung wesentlicher Aspekte der Thematisierung LIB scheinen diese zwei Verfahren klar zu grobe Themen zu erzeugen, die mehr Aspekte innerhalb der identifizierten Themenstrukturen aufnehmen, als tatsächlich abgebildet werden sollen. Mit einer durchschnittlichen Signifikanz der Korrelationen von 4,758 ist der überwiegende Teil der Themen in Bezug auf die Schlüsselereignisse allerdings valide abgebildet.

Innerhalb der Kreuzkorrelationsfolgen kann festgestellt werden, welche „lags“ jeweils zu einer optimalen Korrelation führen. Eine interessante Beobachtung lässt sich machen, wenn für die jeweiligen Korpora ein Durchschnitt für die optimalen „lags“ der Zeiträume erstellt wird. So kann gezeigt werden, dass die Wiedergabe der Ereignisse im Korpus SZ und TAZ um ca. 2 Tage und im Korpus GUARDIAN um 0 Tage verschoben ist. Dies ist plausibel, da die Texte aus dem Korpus GUARDIAN aus einem Online-Nachrichtenportal entnommen sind und demnach ohne Produktionszeit und ohne weitere Verzögerung veröffentlicht werden. Die 2 Tage Verzögerung bei den gedruckten Ausgaben der SZ und TAZ sind durch Produktionszeiten und ausgabenfreie Wochenenden plausibel. Die durchschnittlichen Verzögerungen betragen im Detail 2.183 (TAZ), 1.983 (SZ), 0.05 (GUARDIAN), 2.25 (TAZ_BASE:) und 1.8 (SZ_BASE).

Um die Abbildung der Schlüsselereignisse durch die semi-automatischen Prozesse sichtbar zu machen, werden an dieser Stelle die guten und schlechten Übereinstimmungen in mehreren Grafiken gegenübergestellt. Die Darstellung wird auf aussagekräftige Abbildungen beschränkt, um für die Erläuterung wichtige Beobachtungen zu zeigen. In den Grafiken selbst sind die Schlüsselereignisse noch einmal verzeichnet, sodass die Resonanz in den einzelnen Textquellen gut zu erkennen ist.

Für die Zeitreihen innerhalb der Korpora ist $P^D_{k,t}$ die Größe, die für Analysten am interessantesten ist, da die Dokumentzählungen zeigen, wie viele Artikel ein Thema mit einem ausreichenden Anteil übernehmen. Die beste Übereinstimmung mit der

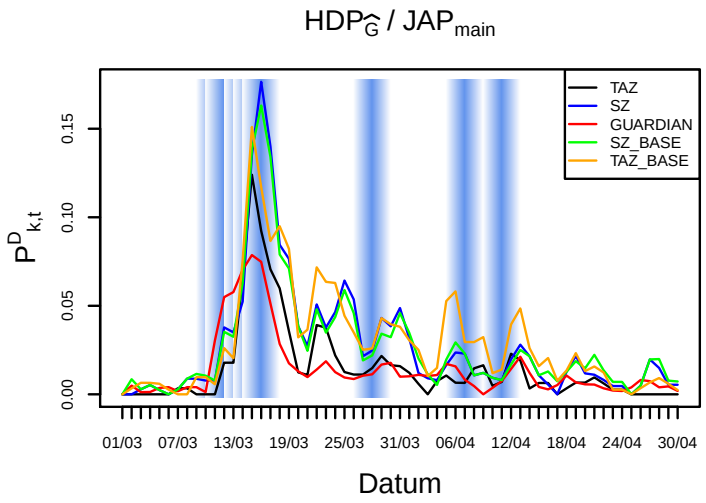

Abbildung 4.11: $P^D_{k,t}$ erzeugt aus HDP$_{\hat{G}}$ der Thematisierung JAP$_{main}$ für alle Korpora.

künstlichen Zeitreihe für das Thema JAP liefert das Verfahren HDP$_{\hat{G}}$. Hier zeigt das eigentliche Hauptthema (JAP$_{main}$) die beste Übereinstimmung mit den Schlüsselereignissen. In Abbildung 4.11 wird die Zeitreihe $P^D_{k,t}$, dieses Themas und Verfahrens in allen untersuchten Korpora dargestellt und die Korrelation der Schlüsselereignisse mit den Ausschlägen ist gut zu erkennen. Es sind Unterschiede im Verlauf in den einzelnen Korpora zu erkennen. Jedoch stimmen der Trend der Kurve und die Ausschläge anhand externer Ereignisse gut überein. Eine weitere Beobachtung ist, dass die Ausschläge im Korpus TAZ_BASE beim Vorhandensein von Schlüsselereignissen deutlicher abzulesen ist, obwohl der Korrelationswert von 0.501 geringer ist, als beispielsweise für den GUARDIAN mit 0.709. Es zeigt sich, dass das Maß der Korrelation nicht mit der Aussagekraft der Verläufe gleichzusetzen ist. Da die künstlichen Zeitreihen an vielen Stellen 0 gesetzt sind, werden Verläufe beim Vergleich bestraft, die in diesen Bereichen dennoch eine starke Dynamik der Intensität aufweisen. Dies ist allerdings nicht von Nachteil, solange die Schlüsselereignisse korrekt abgebildet sind, wie der Unabhängigkeitstest mit der Korrelation gezeigt hat. Für die Thematisierung LIB korrespondieren die Intensitätsverläufe $P^D_{k,t}$, die durch das Verfahren LDA$_{0.5}$ erzeugt werden, am besten mit den definierten Schlüsselereignissen. In der Evaluierung der Zeitreihen hatte sich gezeigt, dass bei den globalen Modellen mehrere Themen erzeugt werden, wobei das Hauptthema wesentliche Schlüsselereignisse nicht abbildet. Bei der Verkettung der Modelle aus den Zeitscheiben zeigt sich aber, dass die Einbeziehung kleinerer Dokumentmengen diese Themen nicht trennt, da zu wenige Dokumente zum Libyenkonflikt in einer Zeitscheibe vorhanden sind, um die

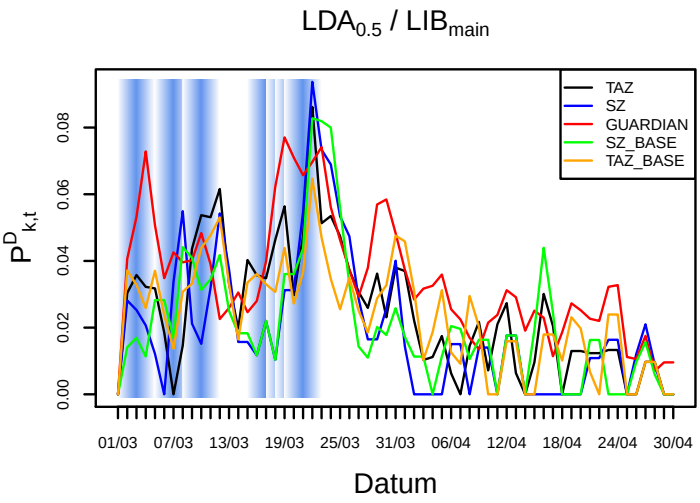

Abbildung 4.12: $P^D_{k,t}$ erzeugt aus $\text{LDA}_{0.5}$ der Thematisierung LIB_{main} für alle Korpora.

zwei Themenstränge voneinander zu trennen. In diesem Fall kann so zuverlässiger identifiziert werden, welche Dokumente einen Bezug haben. Zusätzlich ist interessant, wie der Anteil der eigentlichen Themen am Korpus strukturiert ist. Da die Topic-Modelle nur Teile von Dokumenten zu einem Thema zuordnen können, nämlich über die Wortanteile am Dokument, ist es interessant, die Verläufe der Themenanteile am Korpus $P^W_{k,t}$ den Dokumentanteilen gegenüberzustellen. In Abbildung 4.13 ist der Verlauf der Thematisierung JAP_{main}, entnommen aus den gleichen Verfahren aus Abbildung 4.11, für den Korpus SZ abgebildet. In der Grafik sind die Verläufe $P^D_{k,t}$ und $P^W_{k,t}$ gegenübergestellt. Zwischen beiden Verläufen sind keine großen Unterschiede zu erkennen. In Fällen, wo der Anteil des Themas an den Dokumenten nicht ausreicht, ist aber zu erkennen, dass die Zuordnung der Wörter bzw. Dokumentanteile dennoch kleine Ausschläge erzeugen, die aber für die Analyse von Themenverläufen und Themenphasen anhand von Themeneigenschaften nicht relevant sind. Wesentlich für die Unterscheidung der vorgestellten Verfahren und Parametrisierungen ist die Granularität der erzeugten Themen. Je umfassender und gröber eine Themendefinition ist, desto mehr Dokumente und Wörter können einem Thema zugeordnet werden. Je enger eine Themendefinition ist, desto weniger Dokumente können zugeordnet werden. Der Vergleich der Clustermethode TDT in grafischer Form bietet sich an, um dieses Verfahren bezüglich der Granularität einzuordnen. Im Fall der Anwendung der Verfahren $\text{HDP}_{\hat{G}}$, $\text{HDP}_{\check{G}}$, $\text{HDP}_{0.5}$, $\text{HDP}_{0.6}$ und TDT, dargestellt in Abbildung 4.14 und 4.15 ist dieser Effekt gut zu erkennen. Je feiner die Themen aufgelöst sind, desto weniger Dokumente haben einen ausreichenden Anteil dieser Themen. Besonders der

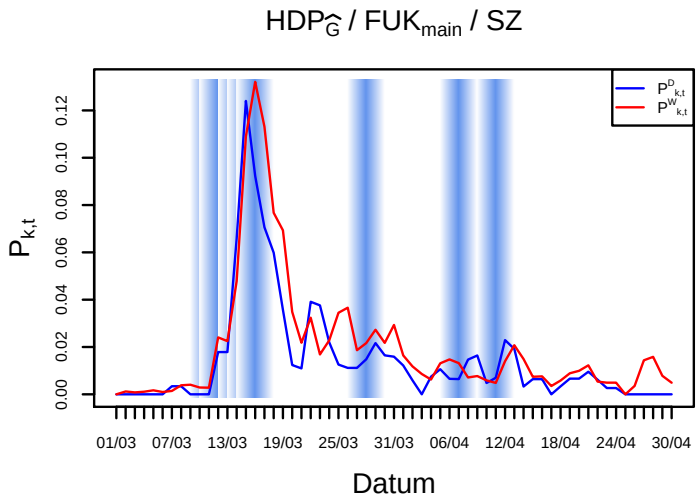

Abbildung 4.13: Darstellung der Veräufe $P^D_{k,t}$ und $P^W_{k,t}$ erzeugt aus HDP$_{\hat{G}}$ der Thematisierung JAP$_{main}$ für den Korpus SZ.

Unterschied der Dokumentzuordnungen bei den globalen Topic-Modellen, wobei den groben Themenberechnungen HDP$_{\check{G}}$ und LDA$_{\check{G}}$ viel mehr Dokumente zugeordnet werden, ist gut sichtbar. In der Abbildung 4.15 ist noch ein weiterer Effekt zu erkennen. Die Abbildung zeigt eine Zeitreihe, die im Test nicht mit der künstlichen Zeitreihe der Schlüsselereignisse korreliert. Hier ist das Modell LDA$_{\hat{G}}$ zu fein aufgelöst und die Selektion des Hauptthemas resultiert in Dokumenten, die eine Thematisierung der internationalen Ereignisse nicht aufnimmt. Somit repräsentiert die Verlaufskurve nicht die Resonanz auf externe Ereignisse. Hier würde der Verlauf der Thematisierung LIB$_{all}$ realistischer sein. Das Verfahren TDT liefert anhand des fest eingestellten Parameters wiederum nur eine mögliche Auflösung der Themen, indem einfach Dokumente geclustert werden. Diese repräsentiert die Hauptthemen aber gut und sie korrespondieren sehr gut mit den Schlüsselereignissen. Die Auswahl der Dokumente ist restriktiver als ein grobes globales Topic-Modell (Bsp. HDP$_{\check{G}}$), jedoch werde einige Dokumente mehr aufgenommen, als bei einem detaillierten globalen Topic-Modell (Bsp. HDP$_{\hat{G}}$). Da die Auflösung bei diesem Verfahren nicht effektiv beeinflusst werden kann, muss das Verfahren im Einzelfall im Hinblick auf Thema und Korpus auf Eignung geprüft werden.

4.3.3 Zwischenfazit

Die Diskussion der Themenintensität zeigt, dass die Verfahren und unterschiedliche Parametrisierungen zu unterschiedlichen aber vergleichbaren Ergebnissen führen. Im

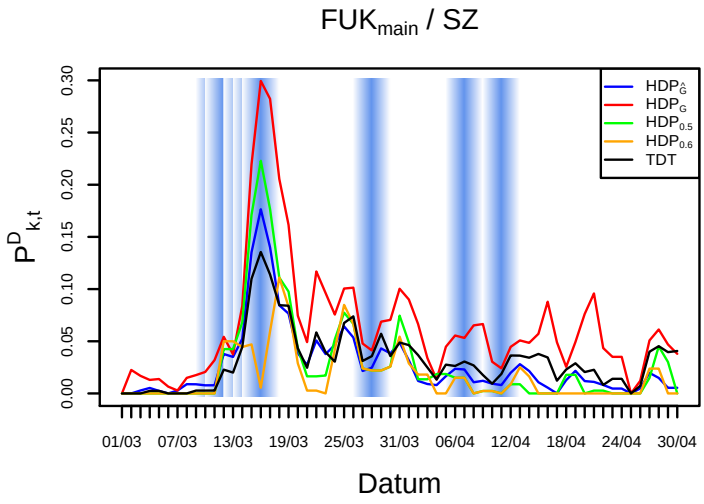

Abbildung 4.14: Darstellung der Verläufe $P^D_{k,t}$ erzeugt aus $\text{HDP}_{\hat{G}}$, $\text{HDP}_{\tilde{G}}$, $\text{HDP}_{0.5}$, $\text{HDP}_{0.6}$ und TDT der Thematisierung JAP_{main} für den Korpus SZ.

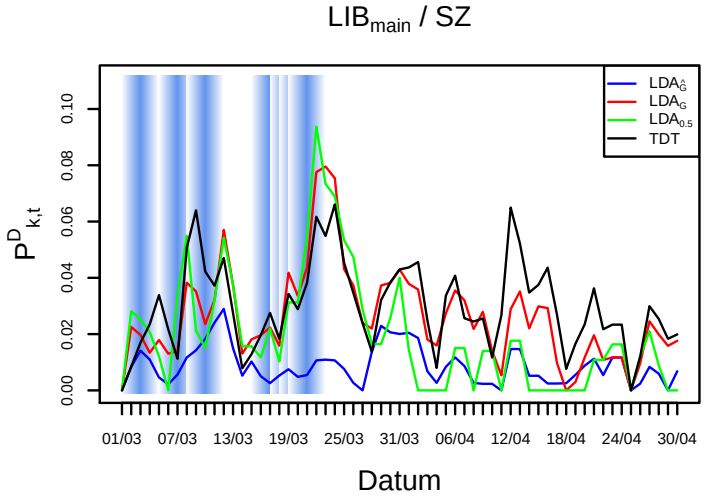

Abbildung 4.15: Darstellung der Verläufe $P^D_{k,t}$ erzeugt aus $\text{LDA}_{\hat{G}}$, $\text{LDA}_{\tilde{G}}$, $\text{LDA}_{0.5}$, $\text{LDA}_{0.6}$ und LDA der Thematisierung LIB_{main} für den Korpus SZ.

Allgemeinen kann die Aussage getroffen werden, dass die von den Verfahren generierten Themen, bis auf wenige zu begründende Ausnahmen, in Abhängigkeit zu externen Schlüsselereignissen stehen. Die Messung der Themenintensitäten kann deshalb als valide angesehen werden. Die Evaluierung zeigt eine Perspektive für das methodische Vorgehen bei semi-automatischen Themenanalysen. Denn bei der Arbeit an neuen Themen (prospektiv) oder der Auswertung bereits bekannter Thematisierungen (retrospektiv), können die hier gezeigten Evaluierungen zum Nachweis der Verfahrensvalidität und -reliabilität eingesetzt werden. Die begleitende Evaluierung kann in zwei Bestandteile eingeteilt werden:

- **Validität:** Wenn es bekannte Einflüsse oder Schlüsselereignisse einer Thematisierung gibt, können diese genutzt werden, um mit Korrelationsverfahren oder "Interupted time series analysis„, deren Einfluss auf eine gemessene Themenintensität zu testen. Ist dieser Zusammenhang nicht gegeben, muss an der Validität gezweifelt werden und genau geprüft werden, ob die Textauswahl, die Wahl des Verfahrens, die Wahl der Verfahrensparameter oder die manuelle Selektion der Ergebnisse korrekt vorgenommen wurden.

- **Reliabilität:** Die Inferenzverfahren für die Topic-Modelle basieren auf Zufallsprozessen. Es ist deswegen immer eine Unsicherheit bzgl. der Themen und deren Wortzusammensetzung zu vermuten. Dadurch werden die Zusammensetzung der Themen und die interpretierbaren semantischen Zusammenhänge nicht bei jedem Durchlauf des Verfahrens stabil bleiben. Jedoch betrifft dies nur einen Teil der berechneten Themen und signifikante thematische Zusammenhänge werden bei mehrmaliger Messung immer abgebildet. Um die Reliabilität des eingesetzten Topic-Modells zu testen, muss eine Testmenge aus dem analysierten Korpus extrahiert werden. Diese Testmenge wird mit identischen Parametern mehrfach berechnet. Die inhaltliche Übereinstimmung der Themen, innerhalb einer solchen Mehrfachmessung, kann als Überprüfung und Nachweis der Reliabilität für ein Topic-Modell an einem ausgewählten Korpus genutzt werden.

Die Feststellung, dass valide Ergebnisse bzgl. themeninterner Schlüsselereignisse erzeugt werden zeigt, dass die vorgeschlagenen semi-automatische Verfahren zur Themenanalyse in der Praxis für die empirische Inhaltsforschung, und Themenanalysen im Speziellen, anwendbar sind. Dadurch kann die von Kolb (2005) beschriebene Unterteilung in Themenphasen, die Identifikation von Nachrichtenfaktoren, die Definition von Schlüsselereignissen und die linguistische Abbildung der Themenbestandteile durch diese Verfahren valide erfolgen. Hinsichtlich der Themengranularität liefern Topic-

Modelle mehr Flexibilität als ein clusterndes Verfahren. Eine grundsätzliche Eignung der Zeitreihen zur Extraktion von Schlüsselereignissen oder Thematisierungen ist aber bei allen Verfahren gegeben. Da das TDT-Verfahren in den Veröffentlichungen (Allan u. a., 1998, 2005) nur für Nachrichtentexte und stark ereignisorientierte Textquellen evaluiert wird, ist die Eignung für andere Textsorten weiter zu untersuchen. Durch die Topic-Modelle ist es möglich, mehrere Themen in Dokumenten zu finden, um so Nebenthemen und Hauptthemen zu trennen. Durch das dokumentorientierte Clustering bei TDT ist dies nur über komplette Dokumente möglich. Themenmischungen innerhalb der Dokumente können nicht aufgefunden werden.

4.4 Wort- und Akteurshäufigkeiten in Themen

Bei der Berichterstattung über Schlüsselereignisse oder einer fortlaufenden Thematisierung spielen verschiedene Terme eine besondere Rolle. Innerhalb eines Themas übernehmen diese Wörter Eigenschaften oder eine Erklärungsfunktion für ein Thema. Diese Wörter werden Schlüsselbegriff genannt (Rauchenzauner, 2008, vgl. S. 39). Es zeigt sich, dass es durchaus interessant ist, die Zusammensetzung eines Themas genauer zu betrachten, um wichtige Begriffe und Bedeutungen zu analysieren. Weiterhin ist es interessant, die Bedeutung oder Nutzung eines Wortes in einem definierten Zeitraum und Kontext zu untersuchen. Dies kann nützlich für viele Teilbereiche einer Inhaltsanalyse sein. Einerseits kann eine Untersuchung der Wortverwendung genutzt werden, um Schlüsselbegriffe einer Thematisierung erkennen zu können. Andererseits kann untersucht werden, ob eine journalistische Publikation unter bestimmten Gesichtspunkten über ein Thema berichtet. Mögliche Fragestellungen wären beispielsweise die Frage nach Termen, die mit Gefahren assoziiert werden oder Terme, die über negative Auswirkungen einer Entwicklung berichten. Die Quantität der genannten Personen oder Organisationen, die innerhalb einer Thematisierung genannt werden, ist interessant, da die Nachrichtenwerte über die Nachrichtenfaktoren, Personalisierung, persönlichen Einfluss, politische Nähe oder Prominenz beeinflusst werden. Vorstellbar ist auch, die Menge der in den Texten enthaltenen Personen in wissenschaftliche Akteure und politische Akteure zu zerlegen, sodass durch eine solche Untersuchung Themenphasen und die Themenreife beurteilt werden kann. In Abschnitt 3.4 werden die Kennzahlen 3.17, 3.18, 3.19 und 3.20 definiert. Diese Kennzahlen erlauben, den Anteil eines Wortes an einer Thematisierung zu bestimmen. Die Berechnung kann normiert werden. Im Fall der Betrachtung von Zeitscheiben, wird die Normierung über die relevanten Wortmengen der jeweiligen Zeitscheibe bestimmt. In diesem Fall lässt sich so eine Zeitreihe bilden, über die die Benutzung eines Wortes diachron analysiert

werden kann. Es ist möglich zu beobachten, welchen Anteil ein Wort an einer Thematisierung einnimmt und ob es einen Trend in der Verwendung eines Wortes gibt. Wie schon im Abschnitt 3.4 erläutert wird, können mehrere Worte als ein Konzept zusammengefasst werden und dieses Konzept kann hinsichtlich der Häufigkeit untersucht werden. In den Thematisierungen JAP können beispielsweise alle Wörter zusammengefasst werden, die die Zeichenkette „strahl“ (Strahlung, Strahlenschutz, strahlendes, Strahlendosis etc.) oder „evaku“ (Evakuierung, Evakuierten, evakuieren, etc.) enthalten, um einen Einblick zu bekommen, welchen Anteil die Referenz auf Gefahren durch Strahlung und den Schutz der Bevölkerung an der Thematisierung haben. Innerhalb der Thematisierung LIB wäre die Zeichenkette „flücht“ (Flüchling, flüchten) interessant, um die Zusammenhänge der durchgeführten Luftschläge zu einer vermehrten Flüchtlingsaktivität zu untersuchen. Die Zusammenfassung unterschiedlicher Terme kann durchaus auch für Eigennamen in den untersuchten Themen stattfinden. Um beispielsweise Personen und Organisationen innerhalb der Themen zu identifizieren, kann eine Eigennamenerkennung als Vorverarbeitung der Texte durchgeführt werden. Bei der Auswertung der Termhäufigkeit innerhalb einer Thematisierung können die Eigennamen zu einem Konzept zusammengefasst werden. Auf diese Weise werden Zeitreihen über Personennennungen oder Organisationsnennungen gebildet. Eine weitere Aufteilung kann erfolgen, wenn beispielsweise Politiker oder private Personen getrennt werden oder nur bestimmte Formen von Organisationen, wie beispielsweise Parteien, innerhalb einer Thematisierung analysiert werden.

4.4.1 Themenabhängige Häufigkeiten von Wörtern

In den folgenden Darstellungen und Erläuterungen werden anhand der Thematisierungen LIB und JAP Querschnittanalysen gezeigt, die Wortverwendungszusammenhänge innerhalb der Themen exemplarisch darstellen. Bei einer ersten Untersuchung der Wortverwendung der Begriffe *erdbeben (earthquake)*,*strahlung (radiation)*,*lebensmittel (food)* und *kernschmelze (meltdown)*, die in Abbildung 4.16 und 4.17 mittels der Kennzahl $S^{W}_{n,k,t}$ dargestellt werden, sind die absoluten Häufigkeiten dieser Terme innerhalb der Thematisierung JAP_{all} dargestellt. Es wird die absolute Häufigkeit genutzt, da die Nutzungsintensität unterschiedlicher Terme in dieser Darstellung gut vergleichbar ist. Die relative Darstellung wäre für den Vergleich unterschiedlicher Korpora mit unterschiedlicher Größe sinnvoll. Es ist an den 2 Beispielen gut zu erkennen, dass im Korpus SZ dem Begriff *erdbeben (earthquake)* mehr Aufmerksamkeit zukommt als im GUARDIAN, wo *strahlung (radiation)* ein größere Rolle im Thema spielt. Ein zweiter interessanter Aspekt ist am Tag des 28. März 2011 erkennbar, an dem eine mögliche

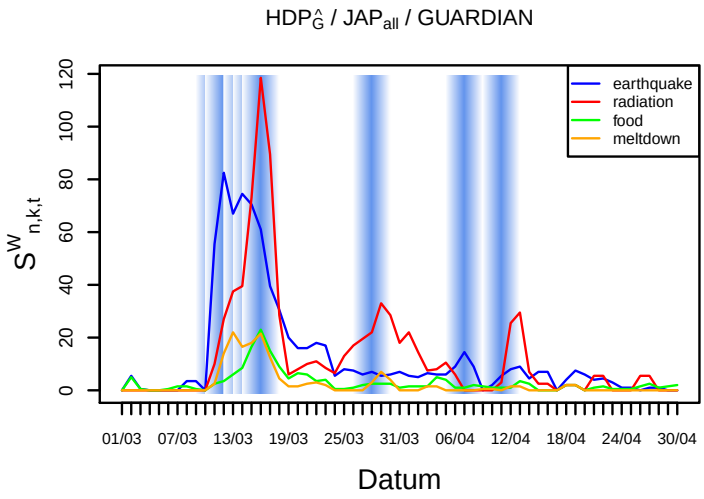

Abbildung 4.16: Darstellung der Verläufe $S^W_{n,k,t}$ erzeugt aus HDP$_{\hat{G}}$ der Thematisierung JAP$_{all}$ für den Korpus GUARDIAN.

Kernschmelze durch die Regierung zugegeben wurde. Auch hier ist der Anteil der Berichte über *strahlung (radiation)* höher, als der Anteil über eine stattgefundene Kernschmelze. Somit wäre weiter zu prüfen, ob sich Berichte in dieser Zeit auf die Folgen konzentrieren oder die Bezüge auf die mögliche Kernschmelze gleichberechtigt abgebildet werden. Diese Fragen kommen erst auf, wenn solche Einzelaspekte in den Themen mit der vorgestellten Einzelwortanalyse sichtbar gemacht und genauer analysiert werden. Im erweiterten Sinne lassen sich Zählungen für einzelne Wörter in den Themen aggregieren. Dies ist beispielsweise schon sinnvoll, wenn es mehrere morphologische Formen eines untersuchten Wortes gibt. Auch die Verwendung synonymer Worte führt dazu, dass unterschiedliche Worte für ein gleichartiges Konzept in den Texten verwendet werden. Aus diesem Grund kann die Querschnittanalyse von Termen auf Wortgruppen erweitert werden. Die Einzelhäufigkeiten der zu aggregierenden Wörter werden für jede Zeitscheibe addiert,$\sum_n S^W_{n,k,t}, n \in C$, und über das Gesamtvokabular der Zeitscheiben $V_{k,t}$ normalisiert. Dabei ist C eine Menge aus Wörtern, die zu einem Konzept zusammengefasst werden sollen. In zwei Beispielen wird an dieser Stelle demonstriert, für welchen Zweck sich diese Art der Zählung nutzen lässt. In den Darstellungen werden alle Terme beachtet, die sich im Kontext eines Themas befinden. Es werden demnach nicht nur Wörter gezählt, die einem Thema zugeordnet sind. Vielmehr werden alle Wörter beachtet, die sich in den Dokumenten befinden, welche einem bestimmten Thema zugeordnet sind. Die Berechnungsvorschrift ist durch die Fallunterscheidung 3.22 auf Seite 108 vorgegeben. So wird bei der Zählung der Konzepte beachtet, welche Inhalte im Kontext eines Themas auf-

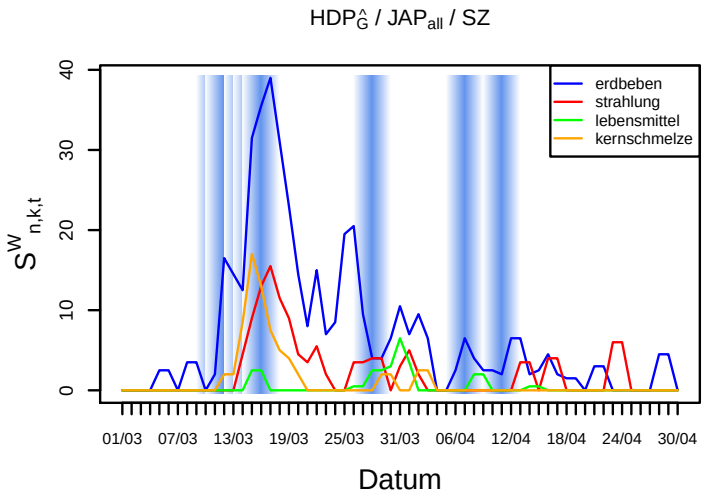

Abbildung 4.17: Darstellung der Verläufe $S^{W}_{n,k,t}$ erzeugt aus HDP$_{\hat{G}}$ der Thematisierung JAP$_{all}$ für den Korpus SZ.

genommen werden. In Abbildung 4.18 ist das Konzept *Evakuierung (Evacuation)* innerhalb der Thematisierung JAP dargestellt. Alle Terme werden zusammengefasst, die irgendetwas mit der Evakuierung von Arbeitern, Hilfskräften oder Privatpersonen zu tun haben. Einige Beispiele sind *evakuierung*, *evakuierungszone*, *evakuierungsradius*, *evakuierungen* oder *evakuieren*. In der Darstellung lässt sich ablesen, dass sich die Referenzen auf Evakuierungen an den Zeitpunkten der Schlüsselereignisse innerhalb der Thematisierung häufen. Durch die relative Darstellung ist jeweils der Anteil des Konzepts im Kontext der Thematisierung darstellt. Das Konzept gewinnt zu den Zeitpunkten der Schlüsselereignisse an Bedeutung. Weiterhin ist zu sehen, dass das Konzept mit der Dauer des Themas einen leicht steigenden Trend hat. Anhand einer solchen Grafik kann also nun die Frage gestellt werden, ob die Evakuierung der Bevölkerung intensiviert wird oder ob die Presseberichte sich diesem Teilaspekt intensiver widmen.

Im zweiten Beispiel wird das Konzept *Flucht (Escape, Getaway, Flee)* in den Korpora GUARDIAN, SZ und TAZ innerhalb des Themas LIB$_{all}$ untersucht. Begriffe wie *flüchtling*, *geflüchteten*, *kriegsflüchtlingen*, *flüchtlingscamps* oder *flüchtlingspolitik* werden zu diesem Konzept zusammengefasst. Es ist erkennbar, dass die Forderung nach internationaler Hilfe (03 März 2011) der Rebellen in Libyen mit einer vermehrten Ansprache des Flüchtlingsproblems in der Presse einhergeht. Dies ist zumindest bei den deutschsprachigen Nachrichtenmedien der Fall. Innerhalb des englischsprachigen Korpus GUARDIAN scheint die Verknüpfung des Konzepts mit der Thematisierung LIB weniger stark ausgeprägt zu sein. Zwischen dem 6. April und dem 15. April

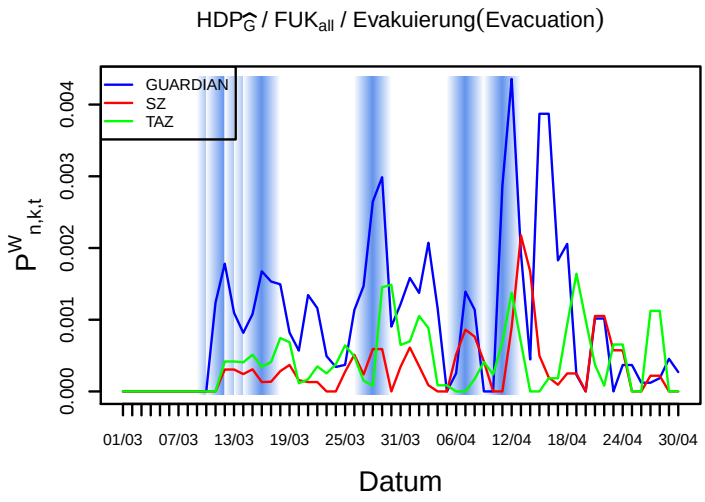

Abbildung 4.18: Darstellung des Verlaufs $P^{W}_{n,k,t}$ für das Konzept „Evakuierung (Evacuation)“ erzeugt aus $\text{HDP}_{\hat{G}}$ der Thematisierung JAP_{all} für die Korpora GUARDIAN, SZ und TAZ.

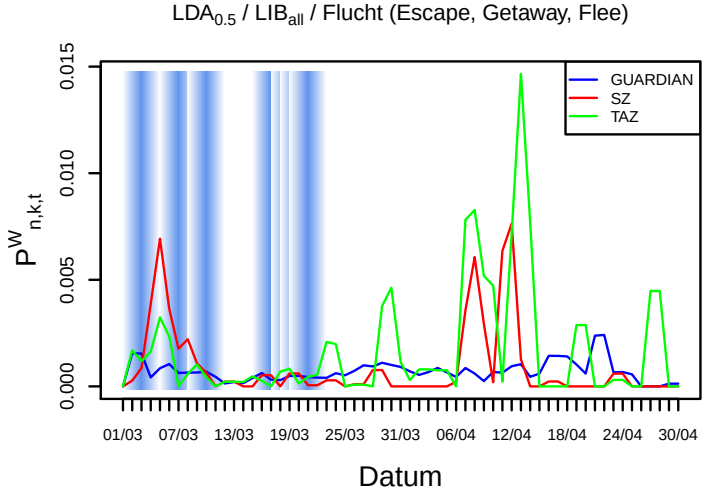

Abbildung 4.19: Darstellung der Verläufe $P^{W}_{n,k,t}$ für das Konzept „Flucht (Escape, Getaway, Flee)“ erzeugt aus $\text{LDA}_{0.5}$ der Thematisierung LIB_{all} für die Korpora GUARDIAN, SZ und TAZ.

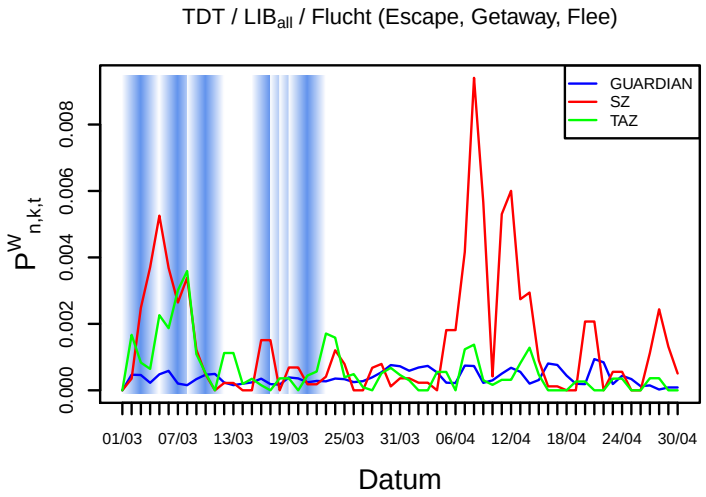

Abbildung 4.20: Darstellung der Verläufe $P^W_{n,k,t}$ für das Konzept „Flucht (Escape, Getaway, Flee)“ erzeugt aus TDT der Thematisierung LIB$_{all}$ für die Korpora GUARDIAN, SZ und TAZ.

scheint das Konzept aber besonders wichtig für das Thema zu sein, obwohl kein typisches Schlüsselereignis für die eigentliche Thematisierung vorliegt. Bei einer genaueren Untersuchung der Dokumentmenge wird deutlich, dass es sich dabei um Berichte über ein gekentertes Flüchtlingsboot handelt. In diesem Fall werden die Geschehnisse aber mit den Ereignissen in Libyen in einen Kontext gesetzt und durch das zugrundeliegende Topic-Modell verknüpft. Anhand dieser explorativen Beobachtung lassen sich auch Hypothesen generieren. Es kann beispielsweise überprüft werden, ob die Flüchtlingsboote, mit denen die Menschen nach Europa kommen, in Libyen starten, da Schleuser es in dieser Zeit besonders einfach haben, Menschen aus afrikanischen Ländern über das Mittelmeer zu bringen. Diese Verknüpfung wird im Korpus GUARDIAN weniger deutlich herausgestellt. Damit stellt sich die Frage nach der unterschiedlichen Darstellung der Flüchtlingsproblematik in unterschiedlichen Ländern, was in weiteren Schritten genauer untersucht werden kann.

Das Beispiel, wie es in Abbildung 4.19 gezeigt wird, basiert auf der Anwendung eines LDA Topic-Modells, welches einzelne Tageszeitscheiben miteinander verknüpft. In Abbildung 4.20 wurde die gleiche Zeitreihe für das das Verfahren TDT erzeugt. Es ist festzustellen, dass der Anstieg des Konzepts *Flucht* um den 6. April 2011 für das Korpus TAZ nicht mehr aufzufinden ist. Zu erklären ist der Unterschied mit der Auflösung, welche die Verfahren jeweils erzeugen. Die TDT arbeitet mit kompletten Dokumenten und ordnet ein Dokument einer „Story“ zu. Dies führt dazu, dass eventuelle Randthemen oder Verknüpfungen, die über die mehrfache Vergabe von Themen pro Dokument hergestellt werden, nicht mehr aufgelöst werden können. Wird

beispielsweise über die Flüchtlingsproblematik berichtet, so wird in diesem Verfahren ein eigenes Cluster definiert, wenn der Bezug zu den Geschehnissen in Libyen nicht ausreichend gegeben ist. In diesem Zusammenhang wäre an dieser Stelle die Überprüfung der Hypothese interessant, ob die Berichte über die Flüchtlingsproblematik in der SZ intensiver mit der Problematik in Libyen verknüpft sind als in den Korpora TAZ und GUARDIAN. Auf diese Weise können Hypothesen und Analysen über Präferenzen unterschiedlicher Medien konzipiert werden, was für die empirische Inhaltsforschung journalistischer Frames oder jounalistischer Kommunikationsabsichten gewinnbringend ist.

4.4.2 Themenabhängige Häufigkeiten von Eigennamen

Die Idee, dass mehrere Wörter als Konzept innerhalb einer Thematisierung untersucht werden können, kann beliebig auf andere Inhalte angewendet werden. Eine wichtige Untersuchung ist beispielsweise die Nennung von Eigennamen oder Organisationen innerhalb einer Thematisierung. Über die Nennung von Eigennamen in den Texten lassen sich zivile, wissenschaftliche oder politische Akteure und Sprecher innerhalb von Texten identifizieren. Damit kann eine Quantifizierung stattfinden, welche dieser Gruppen innerhalb eines Themas eine wichtige Rolle spielen. Dies spielt insbesondere für die Identifikation verschiedener Phasen eines Themas eine wichtige Rolle, da über den Anteil politischer oder wissenschaftlicher Akteure genau bestimmt werden kann, welche Phase eine Thematisierung erreicht hat (Kolb, 2005). So wäre ein Anstieg politischer Akteure ein Zeichen, dass politische Entscheidungsfindungen innerhalb der Thematisierung stattfinden und das Thema bereits etabliert ist.

Über eine „Named Entity Recognition“ können digitale Texte automatisiert nach Nennungen von Eigennamen untersucht werden. Für die hier durchgeführte Analyse wird der „Stanford Named Entity Recognizer“ (Finkel u. a., 2005) genutzt. Dieses Werkzeug bietet Trainingsmodelle für deutsche und englische Nachrichtentexte (Faruqui u. Padó, 2010). Mit diesen Modellen werden alle Texte automatisch mit Informationen über Eingennamen annotiert. Dies ermöglicht die Untersuchung unterschiedlicher Gruppen von Eigennamen innerhalb der Themen über den oben genannten Ansatz. Für die Trennung politischer oder ziviler Akteure müssen die Eigennamen aber noch weiter klassifiziert werden, da bei der automatischen Eigennamenanalyse nur nach Organisationen und Personennamen unterschieden wird. Ein Wörterbuch politischer Akteure, die in einer Thematisierung relevant sind, wäre dafür im Vorfeld zu erstellen. An dieser Stelle wird auf diese Trennung verzichtet. Statt dessen werden die Nen-

nungen von Personen und Organisationen gegenübergestellt, um deren Anteil in den Themen zu vergleichen.

Die folgenden Beispiele sollen anhand der Thematisierung LIB verdeutlichen, wie Nennungen von Person und Organisationen als Wortanteil diachron gemessen werden. Über die Darstellung der Kennzahl $P_{n,k,t}$ sollen Differenzen zwischen unterschiedlichen Kategorien von Eigennamen und Korpora sichtbar werden. In Abbildung 4.21 werden die relativen Häufigkeiten von Organisationen und Personen, die innerhalb der Thematisierung LIB_{all} im Korpus TAZ genannt werden, gegenübergestellt. Anhand der Verläufe ist erkennbar, dass meist mehr Personennamen genannt werden. Der Anstieg der Häufigkeiten korreliert vor allem mit den politischen Reaktionen auf die Ereignisse. Demnach wird die Berichterstattung über die Anerkennung des Übergangsrates durch Frankreich am 10. März 2011 und die politische Entscheidung über die UN-Resolution zwischen dem 17. und 19. März 2011 von einer erhöhten Nennung von Personen und Organisationen begleitet. Die Definition von Themenphasen in Kolb (2005) passt gut zu dieser Beobachtung, wobei Themen in der Phase der politischen Entscheidungen von vielen politischen Akteuren begleitet werden. Die periodischen Lücken entstehen in diesem Fall durch die Pause der Berichterstattung am Wochenende.

Das zweite Beispiel in Abbildung 4.22 vergleicht die Nennungen von Organisationen innerhalb der Thematisierung LIB_{all}. Die Darstellung zeigt, dass der Anteil der Nennungen von Organisationen zwischen den Korpora SZ und TAZ weitestgehend ähnlich verläuft. Zum Zeitpunkt des Beschlusses der UN-Resolution am 17. März 2011 zeigt sich allerdings, dass im Korpus TAZ wesentlich mehr Referenzen auf Organisationen gemacht werden. Mit Hilfe dieser Darstellung werden Zeiträume und Unterschiede aufgezeigt, die durch eine genauere Analyse näher untersucht werden können. Beispielsweise kann überprüft werden, ob die eine oder andere Berichterstattung Meinungen und Aussagen bestimmter Organisationen wiedergibt oder andere journalistische Inhalte im Vordergrund stehen.

4.4.3 Abgrenzung zu Worthäufigkeitsanalysen

Die Erkenntnis, dass bestimmte Terme stellvertretend für eine Thematisierung angesehen werden können, lässt vermuten, dass diese Begriffe eingesetzt werden können, um über eine Schlüsselwort-basierte Suche Dokumente zu selektieren. Innerhalb aller Texte der Korpora SZ und TAZ werden mit Hilfe der regulären Ausdrücke auf Seite 121 alle Dokumente selektiert, die einen der Begriffe enthalten. Es wird gewissermaßen jedes Dokument einer Dokumentmenge D_k zugeordnet, welches einen der definierten

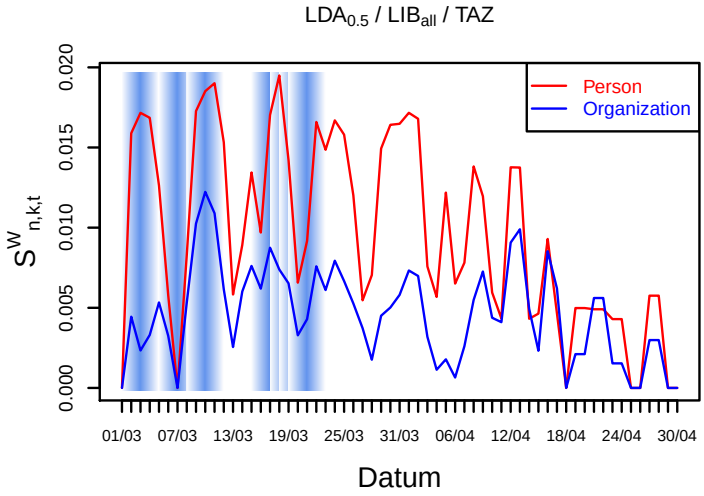

Abbildung 4.21: Darstellung der Verläufe $P^W_{n,k,t}$ für alle aggregierten Nennungen von Personennamen und Organisationen. Die Verläufe wurden aus $\text{LDA}_{0.5}$ der Thematisierung LIB_{all} für den Korpus TAZ hergestellt.

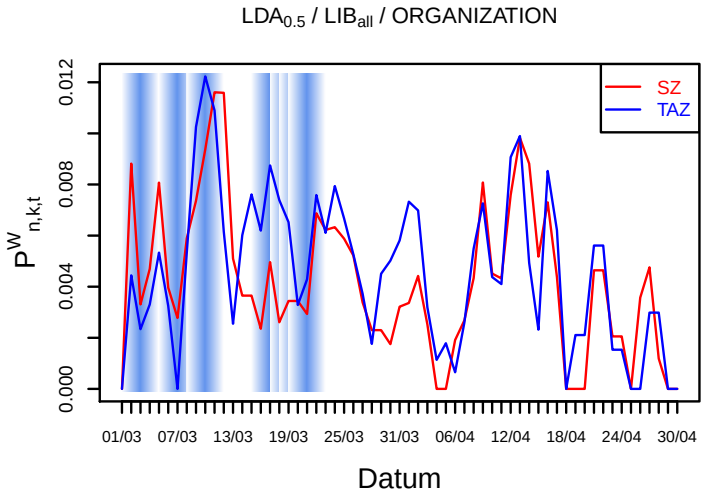

Abbildung 4.22: Darstellung der Verläufe $P^W_{n,k,t}$ für die aggregierten Nennungen von Organisationen. Die Verläufe wurden aus $\text{LDA}_{0.5}$ der Thematisierung LIB_{all} für die Korpora TAZ und SZ erzeugt.

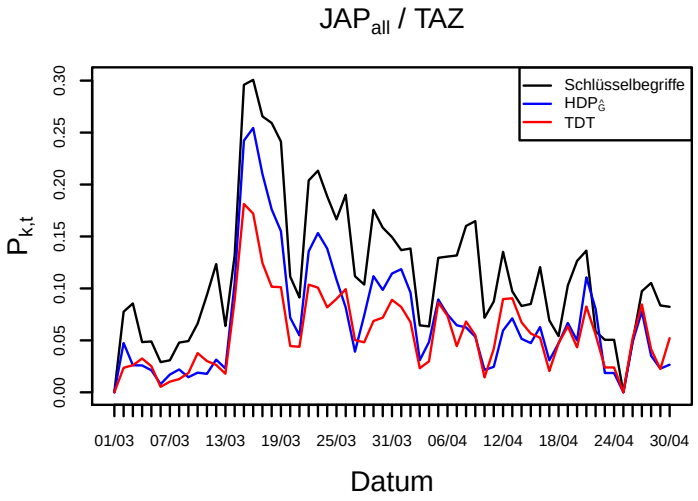

Abbildung 4.23: Darstellung der Verläufe $P_{k,t}$ für die Thematisierung JAP_{all} für den Korpus TAZ. Neben der Dokumentmenge, die aus den Schlüsselbegriffen extrahiert wurde, sind zum Vergleich die Dokumentmengen aus den Verfahren TDT und $\text{HDP}_{\hat{G}}$ dargestellt.

Schlüsselbegriffe im Artikeltext enthält. Für einen exemplarischen Vergleich werden an dieser Stelle die relativen Dokumenthäufigkeiten $P_{k,t}$ der Schlüsselwort-basierten Suche und Beispiele, die mit den vorgestellten Verfahren erstellt werden, in einer Abbildung gemeinsam dargestellt.

In Abbildung 4.23 und 4.24 folgt der Kurvenverlauf der Dokumentmenge, die mit Hilfe der Schlüsselbegriffen extrahiert werden, dem Verlauf der Ergebnisse aus den vorgestellten semi-automatischen Verfahren. Es ist aber zu erkennen, dass bei der Selektion über Schlüsselwörter mehr Dokumente selektiert werden. Die Übereinstimmung der Verläufe zeigt dennoch, dass durch eine valide Auswahl von Schlüsselbegriffen, interpretationsfähige und aussagekräftige Themenintensitäten gemessen werden können.

Dies gilt jedoch nicht für alle Themen. Schon die vorangegangene Darstellung zeigt, dass über Schlüsselbegriffe mehr Dokumente selektiert werden, als bei den vorgestellten Verfahren. Sind die Schlüsselbegriffe zusätzlich mehrdeutig, kann die resultierende Dokumentmenge falsche Dokumente enthalten. Die vorgestellten Verfahren können nicht mit gleicher Qualität von dieser einfachen Methode ersetzt werden. Denn Schlüsselbegriffe einer Thematisierung sind oft nicht einfach erkennbar. Durch Analysen muss erst bestimmt werden, welche Begriffe als Stellvertreter einer Thematisierung gelten können. Dies leisten die vorgestellten Verfahren durch die Abstraktion der Wortvektoren. Schlüsselbegriffe werden erst durch den Einsatz von Verfahren wie Topic-Modellen oder TDT sichtbar. In Fällen, wo die Begriffe bekannt und nicht mehr-

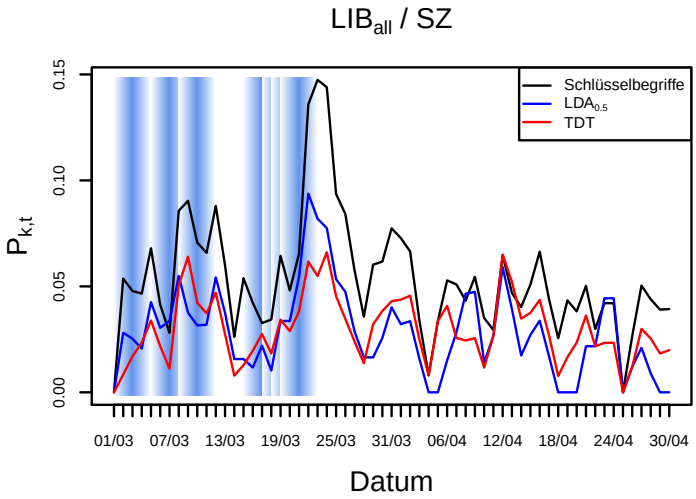

Abbildung 4.24: Darstellung der Verläufe $P_{k,t}$ für die Thematisierung LIB$_{all}$ für den Korpus SZ. Neben der Dokumentmenge, die aus den Schlüsselbegriffen extrahiert wurde, sind zum Vergleich die Dokumentmengen aus den Verfahren TDT und LDA$_{0.5}$ dargestellt.

deutig sind, kann das Selektieren durch Schlüsselbegriffe dennoch schnell und valide eingesetzt werden. Diese Vorgehensweise ist sehr gut an die Themendefinitionen der kategorialen Abgrenzung und der Bezugstheorie anschließbar, da die Schlüsselbegriffe als eine Aufzählung von Nomen verstanden werden können. Für eine Analyse müssen die Begriffe in der Betrachtung der Termhäufigkeiten im Vorfeld bekannt sein, weshalb das Verfahren bei wenig erforschten Themen oder Korpora an seine Grenzen stößt. Somit bieten unüberwachte Verfahren zur Identifikation der Themenstrukturen und -begriffen methodisch und technisch wesentliche Vorteile. Eine einfache Analyse von Termhäufigkeiten und Schlüsselbegriffen kann die Potentiale, Stärken und Vorteile dieser Verfahren nicht erreichen.

4.4.4 Zwischenfazit

Die Analyse des Vokabulars innerhalb einer Thematisierung zeigt verschiedene Perspektiven für den Einsatz in Themenanalysen. Die quantitative Beobachtung von Häufigkeiten bestimmter Themen wird um eine qualitative Komponente erweitert. Werden reine Themenstrukturen betrachtet, so ergibt sich zwar ein Bild, welche Aufmerksamkeit den Themen innerhalb verschiedener Korpora zukommt. Es lassen sich Phasen und Ereignisse ablesen und die Konjunktur von Themen lässt sich verfolgen. Die diachrone Analyse des Vokabulars von Themen erlaubt aber zusätzliche Aussagen über Präferenzen oder Perspektiven, die in unterschiedlichen Textquellen eingenommen werden. Dabei sind nicht die absoluten oder relativen Häufigkeiten eines Wortes

ausschlaggebend für eine Beobachtung. Vielmehr muss ein Begriff oder ein Konzept mit anderen Begriffen verglichen werden, um zu beurteilen, welcher inhaltlicher Rahmen durch eine Zeitung oder einen Journalisten gesetzt wird. Nur über den Vergleich mit anderen Worten und Konzepten oder mit anderen Korpora ist es möglich zu beurteilen, ob Begriffe oder Eigennamen intensiver genutzt werden und als Merkmal einer Thematisierung in einem Korpus interpretiert werden können. So lassen sich Fragen nach Tendenzen, Themenphasen, Schlüsselbegriffen, Schlüsselpersonen oder Akteursstrukturen in Themen stellen. Dies erweitert die Möglichkeit der rein quantitativen Themenbetrachtung um eine weitere Interpretationsebene, die hilfreich für empirische Untersuchungen, vor dem Hintergrund theoretischer Forschungsansätze wie Agenda Setting oder Gatekeeping, ist. Weiterhin erlaubt die detaillierte Analyse des Vokabulars, Nachrichtenfaktoren nachträglich auszuwerten. Eine Analyse, welche Begriffe bei einer erhöhten Aufmerksamkeit, die einem Thema zukommt, eine besondere Rolle spielen, wird möglich. Die wichtigen Begriffe können kategorisiert werden und verschiedenen Nachrichtenfaktoren zugeordnet werden. So kann neben der Analyse der Dauer und Häufigkeit eines Themas nachvollzogen werden, welche Prominenz, Personalisierung oder politische Nähe eine Thematisierung zu einem Zeitpunkt besitzt. Zusätzlich kann nach Begriffen gesucht werden, die Indikatoren auf Nachrichtenfaktoren wie Konflikt, Kriminalität, Schaden oder Erfolg darstellen.

4.5 Analyse des Aussagegehalts in Themen durch Kookkurrenzanalysen

In Abschnitt 3.3 wird die Kookkurrenzanalyse als mögliches Werkzeug für eine Statistik-basierte Analyse themeninterner Aussagen, Propositionen und Valenzen vorgestellt. Anhand der Dokumentmengen, die einem Thema zugeordnet werden können, sei es durch Topic-Modelle oder clusternde Verfahren wie TDT, können die Wörter innerhalb eines Themas in Beziehung gesetzt werden. Es wird demnach für jedes Wort ein themeninterner Kontext extrahiert, sodass die Bedeutung und die Aussagen, in die ein Wort eingebettet ist, automatisch extrahiert werden können. Für die korrekte Einbettung der Wörter in einen Kontext ist es nötig, die relative Position eines Wortes innerhalb einer Analyseeinheit zu beachten. Das Konzept der Nachbarschaftskookkurrenz wird um einen Betrachtungshorizont erweitert, sodass weiter entfernte Wörter relativ zu einem Wort analysiert werden können. Das Ergebnis einer solchen Analyse soll zeigen, welche Aussagen durch die Wörter innerhalb eines Themas gemacht werden. Dies erweitert die bereits diskutierte automatische

Themenanalyse, die eine globale Zusammenfassung als Liste von Begriffen liefert. Das inhaltliche und qualitative Verstehen der Aussagen in einem Thema wird so möglich. So können Aussagen, an denen Schlüsselbegriffe beteiligt sind, extrahiert werden oder Valenzen und Kontexte sichtbar gemacht werden.

4.5.1 Analyse von Schlüsselbegriffen

Themen werden durch Ereignisse modifiziert. So werden an den Schlüsselereignissen einer Thematisierung neue Begriffe geprägt oder deren Kontext und Bedeutung wird durch Ereignisse modifiziert (Rauchenzauner, 2008, vgl. S. 39). Diese Schlüsselbegriffe, denen durch ein Thema eine besondere Bedeutung zugeschrieben wird, werden im Verlauf einer Thematisierung intensiv genutzt und dienen der Strukturierung der Berichterstattung. Sie werden selbsterklärend eingesetzt und sind als Referenz auf ein Thema unerlässlich, um den thematischen Kontext innerhalb eines Textes vorzugeben. Durch den Einsatz automatischer Verfahren können diese Schlüsselbegriffe sichtbar gemacht werden, da die Aggregation der Themeninhalte in Form von gewichteten Termvektoren verfügbar sind. Die Wörter, die ein hohes Gewicht innerhalb einer Thematisierung aufweisen, können als bedeutungstragende Indikatoren oder Stellvertreter für die Thematisierung gelten. Innerhalb der Wahrscheinlichkeitsverteilung $p(\mathbf{w}|z)$, die durch Topic-Modelle berechnet werden, belegen nur wenige Terme einen Großteil der Wahrscheinlichkeitsmasse. Diese Wörter können als Schlüsselbegriffe identifiziert werden, da sie primär in einem gemeinsamen Kontext oder gemeinsam in Dokumenten genutzt werden.

In einer weiteren exemplarischen Analyse werden zwei Dokumentmengen D_k für die Thematisierungen LIB und JAP isoliert. Auf diesen Dokumentmengen werden zunächst signifikante Kookkurrenzen auf Grundlage des Log-Likelihood Signifikanzmaßes extrahiert. Für die Selektion der Dokumente werden die Korpora SZ und TAZ verwendet. Für die Berechnung der Kookkurrenzen werden die Grundformen der Wörter gebildet, um verschiedene morphologische Varianten eines Wortes zu vereinheitlichen. Für die beiden Thematisierungen werden jeweils die Verfahren zugrunde gelegt, die nach der Korrelationsanalyse die beste Übereinstimmung mit den Schlüsselereignissen aufweisen. Im Fall der Thematisierung LIB ist dies $LDA_{0.5}$ und für das Thema JAP wird das Modell $HDP_{\hat{G}}$ gewählt. Die Analyse wird aus Gründen des Umfangs auf wenige Begriffe begrenzt, die als Schlüsselterme innerhalb der jeweiligen Thematisierungen gelten können. Weiter oben wird beschrieben, dass Schlüsselterme die Wörter sind, welche durch bloße Nennung auf ein bestimmtes Thema hinweisen. Aus den Beschreibungen einer Thematisierung, den Termvektoren, kann das reprä-

sentative Vokabular extrahiert werden, indem alle beteiligten Vektoren der Themen zusammengefasst werden (Wiedemann u. Niekler, 2014). Für die Thematisierung LIB können nun Wörter wie *rebell*, *libyen*, *flugverbotszone*, *nato* oder *einsatz* als wichtige Worte identifiziert werden. Für die Thematisierung JAP wären dies *reaktor*, *radioaktivität*, *strahlung*, *fukushima* oder *katastrophe*. Um den Umfang dieser Analyse auf ein Mindestmaß zu begrenzen, welches den Nutzen des Verfahrens aufzeigt, ohne zu komplex zu werden, konzentriert sich die Analyse auf die Terme *flugverbotszone* und *einsatz* für LIB bzw. *strahlung, radioaktivität* und *fukushima* für JAP. Die Analyse für die Terme *strahlung, radioaktivität* und *flugverbotszone* wird detailliert dargestellt. Die Ergebnisse der Terme *einsatz* und *fukushima* werden verkürzt dargestellt, sodass der Arbeitsablauf ausreichend beschrieben ist und die Ergebnisse hinreichend dokumentieren, was diese Analysestrategie leisten kann.

Die Terme werden zunächst global für alle Dokumente D_k einer Thematisierung analysiert. In einem weiteren Schritt wird untersucht, wie die Bedeutungen oder Aussagen einzelner Wörter durch Schlüsselereignisse modifiziert werden. Es wird jeweils eine Dokumentmenge $D_{k,t-}$ und eine Dokumentmenge $D_{k,t+}$ gebildet, die jeweils den Zustand einer Thematisierung vor und nach einem Schlüsselereignis beschreiben. Dabei wird ebenfalls beachtet, dass die Ereignisse mit ca. 1-2 Tagen Verzögerung in den Zeitungen aufgegriffen werden. Für die Thematisierung JAP wird der Zeitpunkt zwischen dem 13. und 14. März 2011 als Zeitpunkt für ein Schlüsselereignis festgelegt und es wird untersucht, wie die Begriffe durch die Ereignisse in ihrer Bedeutung modifiziert werden. Für die Thematisierung LIB wird der 17. März 2011 als Schlüsselereignis gewählt, um die Bedeutungsmodifikationen zu untersuchen. Die Kookkurrenzen werden mit der Log-Likelihood Signifikanz berechnet. Die Berechnung der Kookkurrenzen basiert auf folgenden Vorverarbeitungsschritten:

- Das gemeinsame Vorkommen, der lokale Kontext, wird nur innerhalb von Sätzen bestimmt.
- Für die Bestimmung werden nur Kombinationen von Wörtern gewählt, die einen Abstand von maximal zwei Wörtern in der Kontexteinheit Satz haben.
- Es werden grundsätzlich nur rechte Satz-Nachbarschaften gezählt, um die Folge von Wörtern im Satz zu erhalten.
- Es werden nur Signifikanzen berechnet, wenn eine Kookkurrenz mindestens zwei mal in den Daten gefunden wird.
- Die Wörter werden vor der Analyse in ihre Grundform überführt.

SZ 01.03.2011 - 31.4.2011	SIG	TAZ 01.03.2011 - 31.4.2011	SIG
flugverbotszone → libyen	138,120	einrichtung → flugverbotszone	169,704
flugverbotszone → einrichten	94,044	flugverbotszone → libyen	104,848
durchsetzung→ flugverbotszone	50,884	durchsetzung → flugverbotszone	76,277
flugverbotszone → schutz	38,048	flugverbotszone → fordern	49,171
flugverbotszone → durchsetzen	35,780	errichtung → flugverbotszone	31,031
liga → flugverbotszone	24,554	un-sicherheitsrat → flugverbotszone	24,344
flugverbotszone → angriff	19,717	flugverbotszone → hinweisen	23,542
		liga → flugverbotszone	20,988
		flugverbotszone → zustimmen	20,151
		schonen → flugverbotszone	18,133
		verhängen → flugverbotszone	18,100

Tabelle 4.6: Signifikante Kookkurrenzen berechnet mit dem Log-Likelihood Signifikanzmaß. Die Dokumente wurden mit dem Verfahren $LDA_{0.5}$ und der Thematisierung LIB_{all} selektiert.

- Die Stopwörter werden aus den Sätzen entfernt.
- Innerhalb der Texte werden Wörter ignoriert, die in mehr als 99% und in weniger als 1% aller Dokumente vorkommen.
- Für die Analysen wird ein POS-Tagging unter Verwendung des Stuttgart Tagset durchgeführt. Die Berechnung wird mit penNLP und einem darin integrierten Modell durchgeführt (openNLP). Für die Berechnung werden nur die Wortarten ADJA, ADJD, ADV, NN, NE, VVFIN, VVIMP, VVINF, VVIZU, VVPP, VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF und VMPP genutzt. So werden nur Eigennamen, Nomen, Adjektive, Adverbien und Verben verarbeitet, da diese Wortarten interpretierbare Aussagen und Bedeutungen in Sätzen realisieren.

In einem ersten Schritt wird die komplette Dokumentmenge D_k der jeweiligen Thematisierungen untersucht. In Tabelle 4.6 sind die Kookkurrenzen 1. Ordnung zum Begriff *flugverbotszone* für die Korpora TAZ und SZ dargestellt. Die Tabelle zeigt die Kookkurrenzen und die analysierte Wortfolge im Text. Die Wortpaare sind nach dem Signifikanzmaß sortiert. In beiden Korpora sind die gerichteten Konkurrenzen *flugverbotszone → libyen*, *flugverbotszone → einrichten*, *einrichtung → flugverbotszone* und *durchsetzung → flugverbotszone* als signifikanteste Kookkurrenzen vorzufinden. Weitere Beispiele sind in Tabelle 4.6 dokumentiert. Durch die zusätzliche Verwendung von regulären Ausdrücken werden die Textquellen nach den extrahierten Wortfolgen gefiltert, die in der Kookkurrenzanalyse als signifikant identifiziert sind. So kann für jede Kookkurrenz eine Menge von Sätzen extrahiert werden, die für die semantische Einordnung und Ableitung von Propositionen für das Thema genutzt werden kann. Als regulärer Ausdruck wird

Datum	Beispiel
Süddeutsche Zeitung (SZ)	
02.03.2011	... rasch eine **Flugverbotszone** über **Libyen** schafft, wird ...
03.03.2011	... Plänen für eine **Flugverbotszone** über **Libyen** gearbeitet ...
04.03.2011	... Interesse an einer **Flugverbotszone** über **Libyen** signalisiert habe ...
09.03.2011	... UN-Mandat, um eine **Flugverbotszone** über **Libyen** einzurichten ...
19.03.2011	... Woche dauern werde, eine **Flugverbotszone** über **Libyen** durchzusetzen ...
21.03.2011	... Politiker, der eine **Flugverbotszone** für **Libyen** forderte ...
21.03.2011	... vor dem Sicherheitsrat für eine **Flugverbotszone** über **Libyen** ausgesprochen ...
die tageszeitung (TAZ)	
01.03.2011	... Etablierung einer **Flugverbotszone** über **Libyen**, um die Zivilbevölkerung ...
14.03.2011	... forderte erneut eine **Flugverbotszone** über **Libyen**; erst ...
15.03.2011	... aufgefordert hatte, eine **Flugverbotszone** über **Libyen** zu verhängen ...
18.03.2011	... Verhängung einer **Flugverbotszone** über **Libyen** zur Abstimmung vorlegen
18.03.2011	... Deutschland und andere, einer **Flugverbotszone** über **Libyen** zuzustimmen ...
19.03.2011	... Umsetzung der **Flugverbotszone** über **Libyen**, außerdem sind Militäraktionen ...

Tabelle 4.7: Anhand von Kookkurrenzen und einem regulären Ausdruck extrahierte Belegstellen. Die jeweiligen Terme der Kookkurrenz sind fett markiert.

```
\bWORD1\W+(?:\w+\W+){0,10}?WORD2\b
```

verwendet. Der Ausdruck prüft, welche Textstellen, in denen die zwei Wortformen aufeinander folgen, existieren, sodass maximal 10 Wortformen zwischen ihnen stehen. Unter deren Verwendung lassen sich Belegstellen aus den Texten für die Kookkurrenz *flugverbotszone* → *libyen* extrahieren und ähnlich einer KWIK-Liste darstellen. In der Praxis kann dies für Textdateien effizient mit dem Kommandozeilenprogramm *grep* durchgeführt werden, wobei der Befehl

```
grep -P -i -o '.{0,20}\bWORD1\W+(?:\w+\W+){0,10}?WORD2\b.{0,20}'
```

jeweils 20 Zeichen vor bzw. nach einer gefundenen Übereinstimmung ausgibt, sofern diese vorhanden sind. Mit dieser Methode extrahierte Belegstellen sind in Tabelle 4.7 dokumentiert. Die Kookkurrenz *flugverbotszone* → *libyen* taucht innerhalb unterschiedlicher Aussagen auf und wird demnach mit unterschiedlichen Absichten in die Kommunikation einbezogen, wie Tabelle 4.7 zeigt. Durch die Analyse solcher Fallbeispiele ist es möglich, die Kontexte und Aussagen, die in der Kookkurrenzanalyse extrahiert werden, einzuordnen. Beispielsweise ergibt sich fast immer die gleiche Präposition zwischen *flugverbotszone* und *libyen*. Die Abfolge kann als feststehende Phrase interpretiert werden. Wird das erweiterte Umfeld dieser Kookkurrenz untersucht (KWIC), so lassen sich andere Kookkurrenzen, die in der automatischen Analyse festgestellt werden, wiederfinden. Durch diese Analyse kann vorerst notiert werden, dass die Aussage einer Flugverbotszone über Libyen in Verbindung mit Termen wie *Plänen*, *schafft*, *Interesse*, *einzurichten*, *durchzusetzen*, *forderte*, *verhängen*,

SZ 01.03.2011 - 31.4.2011	SIG	TAZ 01.03.2011 - 31.4.2011	SIG
menge → radioaktivität	117,516	menge → radioaktivität	117,724
radioaktivität → umwelt	99,052	radioaktivität → austreten	90,648
radioaktivität → austreten	98,330	erhöhen → radioaktivität	67,165
radioaktivität → freisetzen	76,943	austritt → radioaktivität	54,957
radioaktivität → gelingen	57,041	radioaktivität → messen	53,609
austritt → radioaktivität	56,126	ausgetreten → radioaktivität	38,734
erhöhen → radioaktivität	55,895	natürlich → radioaktivität	37,877
freisetzung → radioaktivität	52,197	wolfenbüttel → radioaktivität	34,372
groß → radioaktivität	35,889	radioaktivität → umwelt	32,887
ausgetreten → radioaktivität	34,328	radioaktivität → freisetzen	31,447
erhöhte → radioaktivität	34,328	künstlich → radioaktivität	30,559
hoch → radioaktivität	32,825	austretend → radioaktivität	29,320

Tabelle 4.8: Signifikante Kookkurrenzen berechnet mit dem Log-Likelihood Signifikanzmaß. Die Dokumente wurden mit dem Verfahren $\text{HDP}_{\hat{G}}$ und der Thematisierung JAP_{all} selektiert.

zuzustimmen oder *Umsetzung* genutzt wird. Diese Informationen helfen, detailliert zu verstehen, in welche Kontexte thematische Inhalte eingebettet sind. Tabelle 4.8 zeigt dieses Vorgehen an einem weiteren Beispiel, welches die signifikantesten Kookkurrenzen zum Term *radioaktivität* aufführt. Bereits in der Betrachtung der Tabelle können verschiedene Zusammenhänge abstrahiert werden. So wird einerseits die Freisetzung und der Austritt von Radioaktivität besprochen. Andererseits wird aber eine vergleichende Quantität in Form von Mengen oder einer Erhöhung besprochen, welche sich auf einen Normalzustand bezieht, der eine Veränderung erfährt. In Tabelle 4.9 wurde die Kookkurrenz *erhöhen* → *radioaktivität* mit einem entsprechenden regulären Ausdruck in eine Liste von Beispiele überführt, sodass für die Betonung der Differenz zum Normalzustand verschiedene Kontexte sichtbar werden. Ähnliche Kookkurrenzen sind für das Wort *strahlung* zu finden, weshalb für die Kontextbeispiele das Wort *strahlung* als Synonym für *radioaktivität* gesucht wird. In der beispielhaften Auflistung zeigt sich, dass innerhalb der Thematisierung FUK vorwiegend über erhöhte Strahlungswerte bzw. Radioaktivität gesprochen wird. Dennoch lassen sich Beispiele zeigen, die diesen zu erwartenden Kontext nicht erfüllen. So sind Aussagen zu finden, die sich auf die Reaktion, welche die erhöhte Strahlung betreffen, beziehen. In der möglichen Abstraktion dieser Aussagen müssen diese Kontexte gesondert behandelt werden. Für die Formulierung einer Kategorie macht dieser Unterschied der Situation und der Art der Beschreibung inhaltlich viel aus.

4.5.2 Analyse der Auswirkungen von Schlüsselereignissen

Eine weitere Analyseform für die themeninterne Beurteilung von Aussagen und Konzepten ist die Bedeutungsveränderung, die durch Schlüsselereignisse hervorgerufen

Datum	Beispiel
Süddeutsche Zeitung (SZ)	
12.03.2011	... eine Vorsichtsmaßnahme, **Radioaktivität** sei nicht **ausgetreten**.
15.03.2011	... dann große Mengen **Radioaktivität austreten** könnten...
16.03.2011	... es sei viel **Radioaktivität ausgetreten**.
18.04.2011	... drei Monate lang **Radioaktivität** aus der Ruine **austreten** kann ...
21.04.2011	... nur noch wenig **Radioaktivität austreten**. ...
die tageszeitung (TAZ)	
16.03.2011	... Einschätzung des Insiders **Radioaktivität austreten** können.
23.03.2011	... Tepcos Atomkomplex Kashiwazaki **Radioaktivität ausgetreten**, was Tepco aber ...
25.03.2011	... wäre aus dem Reaktor **Radioaktivität ausgetreten** ...
28.03.2011	... deutlich mehr **Radioaktivität** aus dem AKW **ausgetreten** sei als ...
13.04.2011	... weil sie deutlich **höhere radioaktive** Belastungen erlaubten ...

Tabelle 4.9: Anhand von Kookkurrenzen und einem regulären Ausdruck extrahierte Belegstellen. Die jeweiligen Terme der Kookkurrenz sind fett markiert.

wird. Grundsätzlich definieren Schlüsselereignisse die möglichen Interpretationen von Konzepten und Kontexten in einer Thematisierung neu (Rauchenzauner, 2008, vgl. S. 21). Durch eine kontrastive Analyse unterschiedlicher Zeitpunkte können diese Aspekte sichtbar gemacht werden. Der Umfang der Auswirkungen auf verschiedene themeninterne Konzepte oder Terme kann ausschlaggebend sein, um die Tragweite und den Wert eines Schlüsselereignisses zu beurteilen. Um mögliche Bedeutungsveränderungen an Schlüsselereignissen zu erfassen, müssen zunächst die möglichen Schlüsselereignisse einer Thematisierung gefunden und definiert werden. Im einfachsten Fall können die Zeitpunkte mit Hilfe der Zyklentheorie von Kolb (2005) erfasst werden, indem Phasen mit erhöhter Publikationsdichte aus einem Themenlängsschnitt abgelesen werden. Das folgende Beispiel orientiert sich am Schlüsselereignis „17. März 2011 – Einrichtung der internationalen Flugverbotszone (UN-Resolution 1973)“ – der Thematisierung LIB. In Abbildung 4.12 auf Seite 149 ist ab diesem Zeitpunkt ein deutlicher Anstieg der veröffentlichten Dokumente für das Thema LIB abgebildet, der mit einem weiteren Ereignis, die Durchführung erster international geführter Angriffe, seinen Höhepunkt findet.

Für die Analyse der möglichen Veränderungen, die an diesem Zeitpunkt stattgefunden haben, wird die Dokumentmenge zur Thematisierung LIB aufgetrennt. Für die Korpora SZ und TAZ werden anhand der Thematisierung LIB_{all}, extrahiert mit dem Verfahren $LDA_{0.5}$, jeweils alle Dokumente $D_{k,t-}$, deren Zeitstempel zwischen dem 14.03.2011 und 17.03.2011 liegt, und alle Dokumente $D_{k,t+}$, deren Zeitstempel zwischen dem 17.03.2011 und 21.03.2011 liegt, zusammengefasst. Die Dokumentmenge $D_{k,t+}$ beinhaltet ebenfalls alle Texte, die am Tag des Schlüsselereignisses veröffent-

licht wurden. In der Dokumentmenge $D_{k,t-}$ sind diese Dokumente nicht enthalten. Es werden demnach jeweils alle Dokumente vier Tage vor und nach dem Ereignis in eine Dokumentmenge überführt, wobei die Dokumente am Tag des Ereignisses in der Dokumentmenge $D_{k,t+}$ enthalten sind.

In Tabelle 4.10 sind exemplarische Kookkurrenzen für den Term *flugverbotszone* in den Korpora SZ und TAZ dargestellt. Die linke Hälfte der Tabelle entspricht jeweils der Dokumentmenge $D_{k,t-}$ und die rechte Hälfte der Menge $D_{k,t+}$. Besonders im Korpus SZ sind deutliche Unterschiede zwischen den zwei Zeitpunkten festzustellen. Während vor dem 17.03.2011 Kookkurrenzen wie *flugverbotszone* → *einrichten*, *flugverbotszone* → *fordern*, *verhängung* → *flugverbotszone* oder *diskussion* → *flugverbotszone* auffindbar sind, können in den Tagen nach dem UN-Beschluss Kontexte wie *durchsetzung* → *flugverbotszone*, *plan* → *flugverbotszone* oder *einsatzbefehl* → *flugverbotszone* extrahiert werden. Das Konzept einer Flugverbotszone wird vor dem Entschluss, und so geben es die Medien wieder, von einigen Seiten eingefordert oder als mögliche Lösung diskutiert. In den Tagen nach dem Entschluss kommen allerdings Hinweise hinzu, die eine Umsetzung, Durchsetzung und Planung einer Flugverbotszone zum Ausdruck bringen. Dieser Wandel bildet also die Transformation einer diskutierten Lösung in eine politische Entscheidung und deren Durchsetzung ab. Diese Betrachtung kann durch die Hinzunahme von militärischen Kontexten erhärtet werden. Diese Art des Vorgehens kann so bei der Beurteilung von themeninternen Neuinterpretationen durch Schlüsselereignisse herangezogen werden.

4.5.3 Zwischenfazit

Es wurde bereits angesprochen, dass die Betrachtung themeninterner Kookkurrenzen zur Abstraktion, Beschreibung und Kategorisierung von Themen gut an die Theorie von Mackeldey (1987) anschließt. Die darin beschriebene Abstraktion, Konstruktion oder Auslassung von Aussagen basiert auf der detaillierten satzinternen Analyse von Propositionen. Diese Sichtweise kann durch die automatischen Verfahren nur ansatzweise dargestellt werden. Getroffene Aussagen müssen wie Atome eines Themas vollständig heraus destilliert werden. Die oben genannten Beispiele zeigen, dass bestimmte Schlüsselbegriffe einer Thematisierung und deren Wortverwendungszusammenhänge durch eine Analyse der Kookkurrenzen sichtbar gemacht werden können. Durch die Darstellung von Belegstellen signifikanter Muster kann deren Einbettung und Entfaltung in ein Thema am Textbeispiel untersucht werden, was den Interpretationsprozess themeninterner Konzepte unterstützt. Es ist aber anzumerken, dass die Zusammenhänge, die innerhalb der Texte durch Kookkurrenzen gemessen wer-

SZ 13.03.2011 - 16.4.2011	SIG	SZ 17.03.2011 - 21.4.2011	SIG
flugverbotszone → libyen	134,270	durchsetzung → flugverbotszone	39,022
flugverbotszone → einrichten	71,724	plan → flugverbotszone	31,863
diskussion → flugverbotszone	22,872	un-sicherheitsrat → flugverbotszone	29,657
verhängung → flugverbotszone	21,444	flugverbotszone → libyen	27,321
skepsis → flugverbotszone	16,043	flugverbotszone → durchsetzen	25,711
liga → flugverbotszone	13,773	einsatzbefehl → flugverbotszone	24,948
möglichkeit → flugverbotszone	12,099	flugverbotszone → gezielt	24,192
großbritannien → flugverbotszone	10,195	durchgesetzt → flugverbotszone	24,093
vereint → flugverbotszone	9,156	flugverbotszone → schutz	23,148
TAZ 13.03.2011 - 16.4.2011	SIG	TAZ 17.03.2011 - 21.4.2011	SIG
einrichtung → flugverbotszone	137,488	flugverbotszone → libyen	60,987
flugverbotszone → libyen	36,079	durchsetzung → flugverbotszone	45,062
durchsetzung → flugverbotszone	31,738	flugverbotszone → hinweisen	24,267
flugverbotszone → fordern	31,738	schonen → flugverbotszone	20,579
großbritannien → flugverbotszone	23,405	einrichtung → flugverbotszone	16,773
flugverbotszone → libanon	20,455	flugverbotszone → fordern	16,773
entscheidung → flugverbotszone	17,516	un-sicherheitsrat → flugverbotszone	16,682
errichtung → flugverbotszone	16,773	verhängen → flugverbotszone	16,039
flugverbotszone → müssen	16,487	flugverbotszone → zivilbevölkerung	12,398

Tabelle 4.10: Vergleich der Kookkurrenzen aus dem Thema LIB_{all} für die Korpora SZ und TAZ. Es werden jeweils Kookkurrenzen aus den Dokumenten 4 Tage vor dem Beschluss der UN-Resolution und aus den Dokumenten 4 Tage nach dem Beschluss genutzt. Die Dokumente wurden mit dem Verfahren $\text{LDA}_{0.5}$ und der Thematisierung LIB_{all} selektiert.

den, keine vollständigen Propositionen darstellen. Es fehlen Informationen darüber, was Objekt, Relation und Eigenschaft in einem Satz ist. Da bei den Kookkurrenzen immer ein Paar von Wörtern betrachtet wird, fehlt immer eine Komponente. Dieser Beschränkung kann entgegengewirkt werden, indem die Kookkurrenzen als Netzwerk von Wörtern dargestellt werden, die untereinander vernetzt sind. So können zu einem Wort mehrere Zusammenhänge dargestellt werden, die eine Ableitung bzw. Interpretation von Propositionen erlauben. Die Information, ob ein Knoten in einem Graph Objekt, Relation oder Eigenschaft einer Proposition ist, muss idealerweise dargestellt werden können. Dazu ist es nötig, einen weiteren Vorverarbeitungsschritt einzuführen. Dependenz-Bäume müssen für alle Sätze erstellt werden, um syntaktische Relationen in den Daten zu annotieren. Nur so ist eine Unterteilung der Satzbestandteile in elementare Strukturen wie Subjekt, Objekt und Verb möglich, um daraus die Objekte, Relationen und Eigenschaften von Propositionen zu erkennen. Die Auflösung von Koreferenzen kann helfen, die Bezüge zu einzelnen Aktanten im Text aufzulösen, was weitere Analysemöglichkeiten im Rahmen der Themenuntersuchungen ermöglichen würde. So kann auch die funktionale Satzperspektive, bei der Sätze in Thema und

etablierung, einrichtung, umsetzung, abstimmung, durchsetzung, autorisieren, aufgefordert, rasch, fordern, plänen, interesse diskussion, geforderte, berät, militätischen, verhängung, für, ablehnung, folgen, eingerichtete, forderung, resolution, autorisierte, kontrolle	**flugverbotszone**	über, in, für	**libyen**	schafft, gearbeitet, signalisiert, einzurichten, eingerichtet, auszurufen, verhängen, verhängt, gefordert, verlangt, zustimmung, zugestimmt, zustimmen, ausgesprochen, geplant, geebnet, beteiligt, überwachen

Tabelle 4.11: Übersichtliche Darstellung paradigmatischer und syntagmatischer Beziehungen, wie sie vor und nach der Nennung der signifikanten Kookkurrenz *flugverbotszone* → *libyen* in den Korpora SZ und TAZ zu finden sind.

Rhema unterteilt werden, eine Anwendung in der automatischen Analyse von Themen finden.

Die Verwendung von Kookkurrenzen, um damit Propositionen und Kontexte zu bestimmen, lässt sich mit der Theorie paradigmatischer und syntagmatischer Beziehungen nach de Saussure verknüpfen. Die wichtigen Konzepte syntagmatischer und paradigmatischer Beziehungen sind ähnlich der Betrachtung von freien Valenzstellen eines Valenzträgers. Wird eine gerichtete Kookkurrenz betrachtet, stellt dies eine syntagmatische Beziehung dar. Die Nutzung des Wortes entspricht einem, dem Signifikanzmaß nach, im Korpus wiederkehrenden Muster. In Tabelle 4.9 und 4.7 wird solch ein Muster, eingebunden in unterschiedliche Kontexte, dargestellt. Es existieren Wörter in den Lücken zwischen den Begriffen und Wörter, die vor oder nach dem Muster stehen. Dies zeigt, dass es unterschiedliche Ersetzungen gibt, die durch die Kookkurrenzanalyse nicht oder nur teilweise extrahiert werden. Die freien Stellen stellen die paradigmatische Beziehung dar, deren Interpretation als Valenzfüller durchaus zulässig ist. Über eine Analyse der paradigmatischen Beziehungen zu einer gerichteten Kookkurrenz können übersichtliche Darstellungen erarbeitet werden, die es erlauben, Bedeutung und Einbettung einer signifikanten Phrase oder Wortbeziehung schnell zu erfassen. Eine pragmatische Vorgehensweise wäre die Extraktion der Kontexte, wie es für die Tabellen 4.9 und 4.7 beschrieben wird. Zusätzlich können die Kontexte oder paradigmatischen Beziehungen aus jedem Beispielkontext durch zusätzliches POS-Tagging isoliert werden, sodass nur Adjektive, Adverbien, Verben oder Substantive aus dem Nachbarschaftskontext selektiert werden. In Tabelle 4.11 wird dies beispielhaft gezeigt. Viele der Wörter werden als Kookkurrenz zu *flugverbotszone* oder *libyen* aufgefunden. Kookkurrenzen wie *durchsetzung* → *flugverbotszone* gelten als syntagmatische Beziehungen innerhalb der Thematisierung LIB. Werden diese

Darstellungen umfassend analysiert, lassen sich Abstraktionen einzelner thematischer Kontexte, im Sinne von van Dijk (1980) oder Mackeldey (1987), unterstützen. Für das vorgestellte Beispiel können Abstraktionen wie „Vorbereitung einer Flugverbotszone in Libyen“, „Umsetzung einer Flugverbotszone in Libyen“ oder „Kontrolle einer Flugverbotszone in Libyen“ ableiten. Es ist nicht nur eine Verzahnung mit der Thementheorie der Makropropositionen möglich. Der Ablauf dieses Prozesses kann besser strukturiert, organisiert und dokumentiert werden, wie es von Früh (2001) gefordert wird. Dies erleichtert den Anschluss an die Themenzyklentheorie von Kolb (2005), da durch semi-automatische Themenanalysen und die Darstellung syntagmatischer und paradigmatischer Beziehungen jede Phase einer Thematisierung systematisch zusammengefasst und untersucht werden kann.

4.6 Zusammenfassung und weitere Analysemöglichkeiten

In diesem Kapitel wird gezeigt, dass der Einsatz von Text-Mining Verfahren viele Perspektiven für semi-automatische Arbeitsabläufe in der Themenanalyse bietet. Durch eine Evaluation kann belegt werden, dass die vorgestellten Verfahren reliable und valide thematische Zusammenhänge in digitalen Textquellen abbilden. Der manuelle Eingriff spielt aber immer eine wesentliche Rolle. Denn nur durch die manuelle Selektion und die Bestimmung geeigneter Themenauflösungen ist es möglich, valide Messungen hinsichtlich einer Fragestellung durchzuführen. Durch die manuellen Eingriffe und Verknüpfungen unterschiedlicher Themen oder Cluster innerhalb der Verfahren können Ungenauigkeiten der Verfahren korrigiert werden. Entscheidend ist deren Fähigkeit, sich mit Thementheorien der Linguistik verbinden zu lassen. Dies ermöglicht es Anwendern, die mit Inhaltsanalysen vertraut sind, in bekannten Mustern und Vorgehensweisen zu denken.

Die Analyse der Ergebnisse ergibt diverse Möglichkeiten der Auswertung. Es ist möglich, einen Korpus durch die Zuordnung von Dokumentmengen D_k zu einer Thematisierung in Untermengen für Themen zu unterteilen. Die Zeitstempel der Dokumente können genutzt werden, um Längsschnitte der Dokumenthäufigkeiten für ein Thema zu erzeugen. Die Überführung eines Korpus in eine solche Darstellung erlaubt den Anschluss an geläufige Arbeitsweisen der empirischen Inhaltsforschung in der Kommunikationswissenschaft. Vor allem wird die einfache und schnelle Analyse von Themenzyklen nach Kolb (2005) möglich. Die Längsschnitte erlauben, dass Phasen einer Thematisierung festgelegt und abgeleitet werden können. Zusätzlich lassen sich durch die Zeitreihen Schlüsselereignisse extrahieren. Diese zeichnen sich durch eine erhöhte Berichterstattungsmenge bzw. Medienaufmerksamkeit aus (Rauchenzauner,

2008; Kolb, 2005). Die Korrelation mit externen Schlüsselereignissen hat in der Evaluierung gezeigt, dass die so produzierten Zeitreihen abhängig von externen Ereignissen sind. Die Betrachtung der Korrelation zeigte, dass die Korrelation mit einer Verzögerung, einem „Lag“, in einigen Korpora maximal ist. Dadurch lässt sich schlussfolgern, dass Informationen eine gewisse Zeit brauchen, um in einem Nachrichtenmedium aufgenommen zu werden. Die in diesem Kapitel gezeigten Beispiele bestätigen mit der Evaluierung die intuitive Vermutung, dass Online-Textnachrichten eine geringere Verzögerung aufweisen, als gedruckte Zeitungen.

Für die detaillierte Beschreibung und Kategorisierung eines Themas wird gezeigt, wie die Inhalte der Themen abstrahiert werden können. Dies muss im Hinblick auf unterschiedliche linguistische Thementheorien anschlussfähig sein. Die Darstellung und Zusammenfassung der Themen wird in den Verfahren LDA, HDP und TDT durch gewichtete Termlisten hergestellt. Hierbei können Anknüpfungspunkte zur Bezugstheorie und zur kategorialen Abgrenzung gefunden werden, da diese zur Beschreibung thematischer Zusammenhänge auf Listen von Wörtern zurückgreifen. Allerdings sind diese Theorien auf die Darstellung von Nomen beschränkt, da sich die Themen darin als Referenz auf Objekte außerhalb des Textes beziehen. Durch die Gewichtung der Terme in den automatischen Themenanalysen kann eine Wertigkeit einzelner Begriffe innerhalb der Beschreibung eingeführt werden. Aufbauend von dieser Betrachtungsweise der Themen können Visualisierungen entworfen werden, die Themen als Wortgruppen darstellen. Die Einbeziehung der Information über Wortgewichte ist hilfreich, um die Relevanz dargestellter Wörter in einer Darstellung zu betonen. So kann die Aufmerksamkeit auf Worte gelenkt werden, die für die Thematisierung zentral sind. Dies ist ebenso für die Analyse von Schlüsselereignissen und Schlüsseltermen und deren Entwicklung eine wichtige Hilfe. Die Zuordnung einzelner Dokumente zu Themen wird durch die Möglichkeiten der Topic-Modelle erweitert. Die Dokumente können in mehrere Themen aufgeteilt werden. Sofern Themen als Fokus definiert und verstanden werden, wobei Themen als Auswahlfunktion oder Faktor über ein Vokabular dargestellt werden, passt diese Art der Modellierung sehr gut zur Theorie. Die Topic-Modelle erlauben einerseits, Themen als eine Verteilung darzustellen und andererseits, Dokumenten mehrere Themen zuzuordnen. Die Themen haben unterschiedliche Anteile an einem Dokument. Dadurch ist es möglich, Dokumente für ein Thema zu finden, die einen bestimmten „Fokus“ auf das Thema haben. Es kann entschieden werden, wie groß der Anteil eines Themas am Dokument sein muss, sodass es als zugehörig ausgewiesen werden kann. Dies ermöglicht eine genaue Festlegung wie thematisch eng ausgewählte Dokumente sein müssen. Dadurch ergeben sich Anwen-

dungen für die explorative Suche, die empirische Beschreibung von Kategorien oder das Dokument-Retrieval nach thematischen Kriterien. Zusätzlich wird eine erweiterte Sichtweise auf die Themenstrukturen realisiert, wenn in Dokumenten, die einem Thema zugeordnet sind, gemeinsam auftretende Wortformen betrachtet werden. Mit Hilfe von signifikanten Kookkurrenzen unter aufeinanderfolgenden Wörtern können syntaktische und paradigmatische Wortzusammenhänge gefunden werden. Die Beziehungen und Einbettungen, die die Wörter einer Thematisierung erfahren, können dadurch sichtbar gemacht werden. Speziell die Einschränkung einer Kookkurrenzanalyse auf Wortarten wie Nomen, Verben, Adjektive und Adverbien erlaubt eine zusätzliche Abstraktionsebene von Themen, die gut an die Beschreibung als Makropropositionen nach van Dijk (1980) anschließt. So kann der Inhalt eines Themas abseits der Dokumenttexte detailliert beschrieben werden. Die Anknüpfungsmöglichkeiten mit linguistischen Theorien zur Beschreibung von Themen eröffnet eine qualitative Dimension der semi-automatischen Themenanalysen, da die zugrundeliegenden Daten und Ergebnisse vor diesem Hintergrund interpretierbar und nachvollziehbar werden.

Aus den Ergebnissen der semi-automatischen Verfahren, insbesondere den Dokumentmengen D_k, lassen sich weitere Anwendungsmöglichkeiten in anderen Teilbereichen der Inhaltsanalyse beschreiben. In einem neueren Instrument, der Konsensanalyse, wird beispielsweise versucht, die Überdeckung von Themen in unterschiedlichen Medien zu untersuchen (Top, 2006). Die Verkettung der Themen durch den Vergleich der Wortvektoren, die in Abschnitt 3.2.4 gezeigt wird, kann dafür eingesetzt werden. Werden die Korpora unterschiedlicher Zeitungen zu verschiedenen Zeiten betrachtet, so kann über das Verfahren geprüft werden, welche Themen inhaltlich einander zugeordnet werden können. Anhand der Messung ist es möglich, die prozentualen Anteile übereinstimmender Themen zu bestimmen. So kann beispielsweise eine Aussage über die einvernehmliche Adaption bestimmter Themen getroffen werden. Die Verkettung muss nämlich nicht notwendigerweise sequenziell sein, sondern kann auch in parallelen Korpora stattfinden.

Für die detaillierte Betrachtung der Themen können durch zusätzliche Auswertungen auch empirische Analysen vor dem Hintergrund der Nachrichtenwertforschung, der Agenda Setting Theorie und des Gatekeeping Ansatzes möglich sein.[13] Bezüglich der Nachrichtenwerttheorie können die Dimensionen Zeit, Dynamik und Valenz durch Zeitreihen und eine Analyse des Vokabulars schon relativ gut beurteilt werden. Die

[13] Die Theorien sind an dieser Stelle nur als Anmerkungen und Anregung für weitere Anwendungen aufgezählt. Eine ausführliche Erläuterung wichtiger Ansätze ist in Schenk (2007) gegeben.

regionale oder persönliche Nähe der Personen und Akteure in den Texten ist über eine weitere Klassifizierung der Eigennamen im Text möglich. So wird die Möglichkeit geschaffen eine Analyse der Akteurs- oder Personenkreise durchzuführen. Noch einmal muss der Nutzen dieser textinternen Informationen für die Themenzyklentheorie von Kolb (2005) hervorgehoben werden, da Akteure aus Politik, Wissenschaft oder Wirtschaft Indikatoren für unterschiedliche Themenphasen darstellen.

Beim so genannten Gatekeeping-Ansatz und der News-Bias-Forschung wird untersucht, welchen Einfluss die Filterfunktion von Redaktionen auf die Berichterstattung hat. Der Gatekeeping-Ansatz konzentriert sich auf die Entscheidungsträger und die News-Bias-Forschung auf die subjektiven Vorlieben der produzierenden Journalisten. Die Ansätze versuchen modellhaft zu konstruieren, wie ganze Themen, aber auch einzelne Aspekte, durch die Redaktionen aus der Berichtsrettung herausgefiltert oder in der Darstellung verändert werden. Wird diese modellhafte Perspektive genutzt, um redaktionelle Unterschiede bei der Berichterstattung zu analysieren, kann die technische Automatisierung hilfreich eingesetzt werden. Zum einen kann durch den Vergleich von Zeitreihen in unterschiedlichen Korpora überprüft werden, welche Ereignisse jeweils zu einer Berichterstattung führen. So können die Präferenzen einzelner Redaktionen untersucht werden. Zum anderen können ähnliche Themen detailliert über Wortvektoren oder Kookkurrenzen verglichen werden, sodass Unterschiede zwischen den Redaktionen sichtbar werden. Mit vergleichenden Analysen können Differenzen einzelner Redaktionen und so deren Filterfunktionen identifiziert werden. Bei der Analyse einzelner Redaktionen muss mit einer Referenz gearbeitet werden, sodass Unterschiede zu einer idealen und vollständigen Berichterstattung festgestellt werden können.

In der Agenda-Setting-Forschung wird versucht, einen Zusammenhang zwischen dem Inhalt der Berichterstattung und der Einstellung der Rezipienten zu unterschiedlichen Sachverhalten herzustellen. Dabei ist es wichtig, wie relevant die Rezipienten einen bestimmten Inhalt finden oder ob eine bestimmte Position vertreten wird. Die semi-automatische Themenanalyse wird relevant, wenn Einstellungen und Interessen, die mit einer Befragung ermittelt werden, durch umfassende Themenanalysen ergänzt werden sollen. Die Einstellungen und Präferenzen der Befragten kann so schnell mit einer tatsächlich stattgefundenen Berichterstattung verglichen werden.

Zusammenfassend lässt dieser kleine Ausblick den Schluss zu, dass die semi-automatische Inhaltsanalyse als grundlegende Technologie innerhalb verschiedener empirischer Untersuchungsansätze genutzt werden kann.

Kapitel 5

Diskussion der Forschungsfragen zu automatisierten Themenanalysen

In Kapitel 2 Abschnitt 2.4 werden aufbauend auf den abstrakten Untersuchungsschwerpunkten aus Kapitel 2 detaillierte Forschungsfragen definiert. Sie hinterfragen die grundlegende Vereinbarkeit von Anforderungen der Inhaltsanalyse, und speziell der Themenanalyse, mit den Ergebnissen aus automatischen Verfahren. Die Verarbeitung und Anwendung der Ergebnisse automatischer Methoden zur Themenanalyse in großen digitalen Korpora bilden einen weiteren Aspekt der Arbeit. In diesem Kapitel sollen diese Fragen abschließend diskutiert und beantwortet werden.

5.1 Grundsätzliche Fragen

Ausgehend von der Frage, wie computergestützte Methoden die Themenanalyse unterstützen können (**F1**)[1], werden zwei Teilfragen an den Anfang der Arbeit gestellt. Zum einen soll geklärt werden, wie die Sichtweisen und die Anforderungen der Inhaltsanalyse in einer Themenanalyse mit computergestützten Methoden abgebildet werden können. Eine Definition ist nötig, wie Themen innerhalb von Textsammlungen analysiert und dargestellt werden können. Die zweite Frage sucht nach Möglichkeiten der Automatisierung in der Themenanalyse. Die folgende Diskussion soll diese Fragen abschließend beantworten.

[1] Die Forschungsfragen sind auf Seite 53 dokumentiert.

5.1.1 Anschlussfähigkeit an die Methodik der Inhaltsanalyse

Die methodischen Randbedingungen der Inhaltsanalyse erfordern, dass quantitativ zu messende Variablen innerhalb eines Kategoriensystems definiert werden. Kategorien stellen in diesem Sinne eine erdachte, qualitative, auf Theorien bezogene und inhaltlich definierte Menge von Klassen dar, denen Inhalte zugeordnet werden können. Entspricht das Kategoriensystem den Forderungen erschöpfend, unabhängig, vollständig, eindimensional und trennscharf zu sein, so kann es verwendet werden, die Existenz der Kategorien in einem Text quantitativ und valide zu bestimmen. Demnach müssen erst qualitative Festlegungen und Definitionen getroffen werden, um quantitativ zu arbeiten. Die inhaltliche Beschreibung der Kategorien und deren mögliche Interpretation hängt von theoretischen Vorannahmen oder Vorkenntnissen ab. Die Interpretation der Kategorien muss gewährleistet sein, damit eine Messung nicht nur eine reine Beschreibung inhaltlicher Strukturen in einem Text ist. Um die Kategorien für eine Themenanalyse möglichst erschöpfend und im Sinne existierender theoretischer Annahmen zur Entfaltung von Themen in Texten zu beschreiben, werden in Abschnitt 2.1.3 unterschiedliche linguistische Theorien zur Zusammenfassung und hinreichenden Beschreibung von Themen vorgestellt. Die vorgeschlagenen automatischen Verfahren produzieren Datenobjekte, die sich auf unterschiedliche Art und Weise mit diesen theoretischen Themendefinitionen verbinden lassen. Aus den untersuchten Methoden der Topic-Modelle, des Clustering mittels TDT und der Kookkurrenzanalyse werden

- Dokumentmengen,
- gewichtete Termvektoren und
- gewichtete Verbindungen (Kookkurrenzen)

zu einzelnen Themen erstellt. Dem gegenüber stehen die Festlegungen der linguistischen Theorien. Diese führen Themen auf

- gewichtete bzw. ungewichtete Nominalgruppen (kategoriale Abgrenzung, Bezugstheorie),
- Faktoren (Fokustheorie),
- Bezugsobjekte (funktionale Satzperspektive) und
- propositionale Inhalte (Makropropositionen) von Sätzen zurück.

Die allgemeinen Definitionen, welche die Themen als einen zentralen Gegenstand im Text definieren, sind für die Anknüpfung an eine kategoriale Beschreibung von Themen nicht geeignet. Der Versuch, den zentralen Gegenstand durch unterschiedliche Konzepte eines Themenkerns zu beschreiben, lässt wiederum die Übertragung auf automatische Methoden zu. Einzig die Sichtweise der funktionalen Satzperspektive ist kaum anwendbar und herstellbar. Die Darstellung eines Textes als thematische Progression hilft, die Details einer Argumentation oder Berichterstattung für eine Analyse sichtbar zu machen. Da die Konzentration auf Sätze und das Fehlen von Abstraktionsregeln aber keine hinreichende und textübergreifende Beschreibung eines Themenkerns zulässt, ist diese Methode für eine Analyse einer Vielzahl von Texten ungeeignet.

Die Darstellung von Themen als gewichtete oder ungewichtete Nominalgruppe kann durch die Verfahren der Topic-Modelle und durch Clustering hergestellt werden. Die Dokumentmengen können durch Anwendung des TF/IDF-Maßes, wie es im TDT-Verfahren zur Anwendung kommt, als Vektoren von Wörtern, die jeweils mit einem bestimmten Gewicht im Dokument vorkommen, dargestellt werden. Über die Bildung von Mittelwertvektoren aus allen Dokumenten, die einem Thema zugeordnet sind, kann ein Thema als sortierte Liste von Wörtern dargestellt werden. Bei der Anwendung von Topic-Modellen werden die Themen durch multinomiale Verteilungen repräsentiert. Eine solche Verteilung kann als gewichtete Auflistung des Vokabulars verstanden werden.[2] Topic-Modelle sind generative Modelle, die auf der Annahme beruhen, dass die Wörter eines Dokuments durch verschiedene, latente Themen erzeugt werden. Diese Annahme zeigt deutliche Parallelen zur Fokustheorie, bei der ein Thema als Auswahlfunktion oder Faktor betrachtet wird.

Da die sortierten Listen zur Themenbeschreibung sehr viele Wörter enthalten können, überschneiden sich einige Wörter innerhalb unterschiedlicher Themen eines Korpus. Die Anforderung, dass Kategoriensysteme trennscharf sein müssen, ist damit nicht erfüllt. Dies liegt daran, dass die Listen auch Wörter und deren Gewicht enthalten, die weniger semantisch sind. So können Wörter, die syntaktische Funktionen haben, in allen Themen auftauchen. Diese Schwäche kann aber relativiert werden, indem tatsächlich nur eine Untermenge der Wörter verwendet wird, die ein hohes Gewicht für eine Thematisierung aufweisen. Solche „Ausschnitte“ liefern unabhängige und trennscharfe Zusammenfassungen, die als Kategorisierung tauglich sind. Allerdings können die Überlappungen der Wörter auch eine Information über

[2] Dieses Vorgehen wird in Abschnitt 3.1.2 auf S. 62 ff. und in Abschnitt 3.2.4 auf S. 88 ff. diskutiert.

die mögliche Verwandtschaft verschiedener Themen sein. Um aus den Listen reine Aufzählungen von Nomen zu erstellen, müssen vor der algorithmischen Verarbeitung der Dokumente Wortarten gefiltert werden. Durch ein vorgelagertes POS-Tagging ist dies technisch möglich. Es muss allerdings genau geschaut werden, ob für ein Thema auch Wörter relevant sind, die keine Nomen darstellen, aber dennoch wichtig für die Kategorisierung eines Themas sind. Meist sind dies Verben oder Adjektive.

Themen werden um deren zentrale Wörter und Schlüsselwörter entfaltet. Die Wörter werden in einen Kontext gesetzt, treten in Interaktion und sind in Handlungen und Modifikationen integriert. Diese Entfaltung wird in der kategorialen Abgrenzung und in der Bezugstheorie nicht beachtet. Die Extraktion von Makropropositionen oder die funktionale Satzperspektive versuchen dagegen die Entfaltung zu abstrahieren. Innerhalb dieser Theorien stehen einem Betrachter demnach mehr Informationen über ein Thema zur Verfügung. Ein Kategoriensystem für die Themen lässt sich detaillierter erstellen. Dadurch werden mehr Anwendungen, Auswertungen und Inferenzen möglich, als bei der reinen kategorialen Beschreibung. Über die Anwendung der Kookkurrenzanalyse können Propositionen und Modifikationen zu den „Schlüsselbegriffen“ einer Thematisierung sichtbar gemacht werden. In Visualisierungen durch Keyword-in-Context Ansichten, Kookkurrenz-Tabellen oder Kookkurrenz-Graphen, lassen sich dokumentübergreifende Nutzungsmuster von Wörtern beschreiben. Dadurch lässt sich ein Blick auf das Vokabular einer Thematisierung entwickeln, der weit über die reine Beschreibung von Wortgruppen hinausgeht.[3] Durch die weiterführende Arbeit mit den Kookkurrenzen, indem diese zu Textaussagen zusammengefasst werden, können Abstraktionen erstellt werden, sodass ein Thema im Sinne der Theorie von Makropropositionen beschrieben werden kann. Dieser Prozess kann nicht vollständig automatisiert werden, da manuelle Festlegungen und Interpretationen eine wesentliche Rolle spielen, solange die Funktionen der Wörter in den Sätzen nicht automatisch klassifiziert wird. Dafür wäre eine Untersuchung nötig, ob durch semantisches Parsing, Koreferenzanalyse und Dependenzanalyse weitere Daten aus den Texten extrahiert werden können, die eine Übersetzung der Kookkurrenzen in eine propositionale Form oder die funktionale Satzperspektive zulassen.

5.1.2 Automatisierung der Inhalts- bzw. Themenanalyse

Die Frage nach der Automatisierung von themenorientierten Inhaltsanalysen lässt sich durch den Einsatz computergestützter Methoden und deren Anschlussfähigkeit an

[3] In Abschnitt 3.3 auf S. 97 ff. und Abschnitt 4.5 auf S. 164 ff. werden diese Anwendungsmöglichkeiten detailliert besprochen

linguistisch motivierte Kategoriensysteme folgendermaßen beantworten. Grundsätzlich bieten die Verfahren die Möglichkeit einer Automatisierung, da große Textquellen effizient in deren thematische Struktur überführt werden können. Letztlich muss aber festgestellt werden, dass es im Arbeitsablauf manuelle Schritte einzuhalten gilt, um die Messungen im Sinne einer Inhaltsanalyse valide durchzuführen. Die manuellen Schritte umfassen

- die Entscheidungen über den Abstraktionsgrad der Verfahren[4],
- die Beurteilung von Themenbezügen oder Abhängigkeiten[5],
- die Identifikation von Schlüsselbegriffen und Schlüsselereignissen,
- die Interpretation der Themeninhalte und
- die Auswertung von Zeitreihen und Themenphasen.

Die Auflistung zeigt, dass die Analyse von digitalen Textkorpora durch computergestützte Methoden nicht vollständig automatisierbar ist. Vielmehr wird die Themenanalyse durch die computergestützten Werkzeuge effizienter. Die Methoden reduzieren einen Textkorpus auf Dokumentmengen, die einen Wortverwendungszusammenhang bestimmter Abstraktion enthalten. Die Verfahren generieren Sichtweisen auf einen Korpus, die ein sogenanntes Distant Reading nach Moretti (2005) zulassen. Anhand dieser reduzierten Informationsmenge ist es möglich, Dokumentkollektionen nach gesuchten Zusammenhängen, also Themen, zu filtern und deren thematische Kategorien und Eigenschaften zu beschreiben. Der Zugriff auf die Dokumente ist effizienter und vollständiger, muss jedoch durch manuelle und qualitative Einschätzungen, die an der Fragestellung der Analysten hängen, spezifiziert und validiert werden.

Den Anforderungen der Kategorisierung und der Festlegung quantitativer Kennzahlen wird der vorgeschlagene Automatisierungsansatz gerecht. Die interpretationsfähigen Wortvektoren und die Größen in Abschnitt 3.4 erlauben die Festlegung unterschiedlicher Skalenniveaus für die Beurteilung der Thematisierungen. Durch die Darstellung der Themen hinsichtlich linguistischer Theorien können Gruppen bestimmter Eigenschaften (Nominalskala) identifiziert werden. Auch die Sortierung

[4] Die Möglichkeiten der Einflussname werden in Abschnitt 3.2 auf S. 69 ff. besprochen. Speziell der Exkurs zur Bedeutung der verdeckten Variablen im LDA Modell zeigt deren Wirkung auf den Abstraktionsgrad.

[5] Die Möglichkeiten Bezüge manuell zu beurteilen, werden in Abschnitt 4.2.1 auf S. 117 ff. besprochen.

von Gruppen oder Werten (Ordinalskala, Intervallskala) ist über die Themenanteile sowohl in synchronen als auch in diachronen Analysen möglich. Die Relationen zwischen den Themen können überprüft werden und die Relevanz der einzelnen Themen oder deren Nachrichtenwerte können verglichen werden. Die Themenanteile werden aber auch einem konkreten Wert (Rationalskala) zugeordnet, sodass die automatische Themenanalyse unterschiedliche Skalenniveaus für qualitative und quantitative Zugriffe auf die Themen herstellen kann.

Der Ablauf einer Themenanalyse kann als technischer und methodischer Ablauf folgendermaßen definiert werden:

1. Import der Textdaten
2. Vorverarbeitung der Textdaten (Texttransformationen, Pruning, Vergabe von Wortnummern)
3. Indizierung der Zeitstempel und Aggregation von Zeitintervallen
4. Bildung von zeitabhängigen Dokumentmengen
5. Übergabe der Dokumente an die Analyseverfahren in Form einer Stapelverarbeitung
6. Durchführung einer explorativen Analyse der Themen zur qualitativen Aufbereitung der Ergebnisse
7. Verarbeitung der Ergebnisse zu Zeitreihen und Kookkurrenzanalysen
8. Analyse, detaillierte Interpretation und Schlussfolgerungen anhand der Themen

Um diese vorgeschlagene methodische Vorgehensweise technisch zu realisieren, wird in Anhang A eine Softwareumgebung vorgeschlagen. Mit Hilfe einer Implementierung, die diesem Beispiel folgt oder ähnliche Funktionalitäten zur Verfügung stellt, sind die gezeigten Methoden-Abläufe vollständig durchführbar.

5.2 Erweiterte Fragen

Aus den Anforderungen und Arbeitsweisen der Themenanalyse ergeben sich weitere Fragen. So muss bei einer Verzahnung mit der Methodik der Inhaltsanalyse geklärt werden, ob die Verfahren generell qualitativer oder quantitativer Natur sind. Weiterhin muss die Validität und Reliabilität der Verfahren erläutert werden. Ein wichtige Rolle spielen die Möglichkeiten der Weiterverarbeitung der Ergebnisse, sodass Auswertungen im Sinne kommunikationswissenschaftlicher Theorien möglich werden.

5.2.1 Qualitative und quantitative Aspekte

Vor dem Hintergrund der Differenzierung zwischen quantitativen und qualitativen Methoden, muss die Frage beantwortet werden, wie sich semi-automatische Verfahren mit Hilfe von Topic-Modellen oder Clusteransätzen in diese Unterscheidung einordnen (**F2**). Dabei muss zwischen der Art der Messungen und der Darstellung der Ergebnisse differenziert werden. Die Diskussion soll aber an dieser Stelle nur darüber geführt werden, welche Perspektiven die Verfahren in dieser Hinsicht bieten. Über die grundsätzliche Verwendung innerhalb quantitativer oder qualitativer Forschungsansätze soll hier nicht weiter referiert werden, da an dieser Stelle lediglich die Eignung der Verfahren für die zwei Ansätze erläutert werden soll.

Das in Kapitel 3 und 4 vorgeschlagene Gerüst für die Erstellung von Themenanalysen umfasst explorative und deskriptive Komponenten. Zum einen können die Ergebnisse der Verfahren in geeigneten Visualisierungen dargestellt werden.[6] Die Akzentuierung wichtiger Worte über die Wortgewichte in den Modellen erlaubt eine inhaltliche Einordnung der gefundenen Themen. Durch die gemeinsame Darstellung ergibt sich ein Kontext, mit dessen Hilfe die Dokumentmengen einem bestimmten Sinn zugeordnet werden können, wie beispielsweise einer nominalen kategorialen Unterscheidung in verschiedene thematische Kategorien. Weiterhin können die Kookkurrenzanalysen innerhalb der Dokumentmengen, die einem Thema zugeordnet sind, genutzt werden, um die Bedeutung von zentralen Begriffen über KWIC Darstellungen oder Kookkurrenzgraphen zugänglich zu machen. Dies alles stellt eine qualitative Komponente des vorgeschlagenen Gerüsts dar. Die thematischen Kategorien werden explorativ erfahrbar. Im Hinblick auf eine Vorstellung oder Theorie, im Sinne der explorativen Kategorienbildung und vor dem Hintergrund linguistischer Theorien können die Themen so interpretiert werden. Auf der anderen Seite ordnen die Verfahren zu jedem Thema eine Menge von Dokumenten zu. Auf diese Art und Weise lassen sich Quantitäten für die Themen angeben. Dokumentmengen, Anteile am Korpus, Wortmengen oder Eigennamenanteile können angegeben werden und als Zeitreihe visualisiert werden, sofern diachrone Daten vorliegen. Diese Möglichkeiten entsprechen einer quantitativen Arbeitsweise. Abbildungen und Einordnungen auf der Ordinal-, Intervall- und Rationalskala sind für die sychrone und diachrone Betrachtung quantitativer Themeneigenschaften möglich. Die komplementäre Verwendung der quantitativen und qualitativen Perspektiven erlaubt neben der Bestimmung

[6] Die Perspektiven für die Darstellung werden in Abschnitt 4.2.2 auf S. 125 ff. gezeigt

der thematischen Struktur in einem Korpus auch die intersubjektive Interpretation manifester Themen.

5.2.2 Deduktive und induktive Charakteristiken

Die Frage, ob das vorgeschlagene Gerüst deduktives oder induktives Arbeiten ermöglicht, kann mit Argumenten für beide Arbeitsweisen beantwortet werden (**F3**). Offensichtlich erlaubten die unüberwachten Verfahren, unbekannte Datenmengen zu explorieren. Neben der Extraktion von Zeitreihen werden qualitative Möglichkeiten beschrieben, die Bedeutung der Themen zu erfassen. Dieses induktive Arbeiten erlaubt, Theorien zu entwickeln, indem von den gegebenen Beobachtungen abgeleitet wird. Mit den Verfahren kann ein Verständnis über eine Textkollektionen entwickelt werden. Jedoch kann auch eine deduktive Arbeitsweise realisiert werden. Vor einer Analyse kann festgelegt werden, ob nur Themen mit einer bestimmten Zusammensetzung von Schlüsselbegriffen und Aussagen untersucht werden sollen. Diese Arbeitsweise ist vor allem sinnvoll, wenn festgelegt werden muss, in welchem Abstraktionsgrad die Themen untersucht werden sollen. Durch die Festlegung der Topic-Modell-Parameter, welche die Verteilung $p(w|z)$ beeinflussen, kann die Abstraktion der Themen durch Probeläufe getestet werden. Das Verhalten der Topic-Modelle bezüglich des Parameters wird in Tabelle 3.7 auf Seite 86 dargestellt. Es kann überprüft werden, ob ein vorher definiertes Verständnis eines Themas in den Daten zu finden ist. Eine theoretisch fundierte Referenz kann dazu dienen, Differenzen in der Darstellung eines Themas zu bestimmen. Über die externe Definition von Begriffen oder Eigennamen lässt sich zudem deduktiv überprüfen, ob deren Verwendung innerhalb einer Thematisierung vorliegt oder nicht. Werden den automatisch erstellten Themen externe Definitionen von Begriffen, Eigennamen und Kontexten gegenübergestellt, die sich an Annahmen oder Theorien orientieren, können extern motivierte Hypothesen überprüft werden. Dies entspricht einem deduktiven methodischen Vorgehen.

5.2.3 Validität und Reliabilität

In Abschnitt 4.3 wird die Reliabilität und die Validität der Verfahren aus Kapitel 3 überprüft. Die Evaluierungen lassen eine Beurteilung der Validität und Relibilität zu (**F4**). Bei der Reliabilität muss grundsätzlich unterschieden werden, welches Verfahren bewertet werden soll. Das clusternde Verfahren (TDT) arbeitet deterministisch und reliabel, wenn die Dokumente in der gleichen Abfolge in zeitscheibenbasierten Dokumentmengen an das Verfahren übergeben werden. Die Abstandsmessungen mit dem Kosinusmaß und die Berechnungen mit TF/IDF-Termgewichten führen immer

zu einem gleichbleibendem Ergebnis. Die Veränderung des Schwellwertes für die Dokumentähnlichkeit ist nicht sinnvoll. Ein anderes Bild ergibt sich bei der Bewertung der Reliabilität der Topic-Modelle. Die Inferenzverfahren basieren auf Zufallsprozessen und sind aus diesen Gründen nicht deterministisch. Allerdings konvergieren die Inferenzalgorithmen immer zu einem lokalem Optimum der zugrundeliegenden Verteilungen. Wird das Kriterium zugrunde gelegt, dass die Verteilungen in unterschiedlichen Durchläufen exakt gleich erzeugt werden, so liegt der Schluss nahe, dass Topic-Modelle nicht reliabel arbeiten. Durch die Ambiguitäten von Wörtern, die nicht für die semantische Bedeutung eines Themas wichtig sind, ist für den Algorithmus nicht sicher entscheidbar, wo diese Wörter zuzuordnen sind. Allerdings haben diese Wörter für ein Thema kaum Bedeutung, da sich deren auftreten auf alle möglichen Themen verteilt. Das zeigt aber, dass diese Wörter nicht brauchbar sind, um Themen eindeutig und unabhängig zu beschreiben. Die Vollständigkeit der Verteilungen muss deshalb aufgegeben werden und die Reliabilität muss anders bestimmt werden. Auf Seite 138 wird gezeigt, dass bei der Reduktion der zu vergleichenden Terme für die Verteilungen, eine Übereinstimmung der getesteten Modelle von 75 - 80 % erreicht wird. Die nicht stabilen Themen sind zudem tatsächlich von kaum interpretierbarem Inhalt. Die Themen, welche eine Interpretation zulassen, werden in dieser Form auch durch wiederholte Anwendung mit übereinstimmender Zusammensetzung der bedeutungstragenden Wörter generiert. Für die Messung von Kookkurrenzen ist die Reliabilität gegeben, da das Verfahren für gleiche Dokumentmengen stabile Ergebnisse produziert. Allerdings wird erst durch die Verarbeitung der Dokumente in den Themenanalysen festgelegt, welche zugrundeliegenden Dokumentmenge für eine Messung genutzt wird. Diese hängt davon ab, welche Themenanteile für ein Dokument zu einem Thema erwartet und festgelegt werden. Die so generierten Dokumentmengen müssen für die nachträgliche Wiederholbarkeit der Messung dokumentiert sein.

Die Validität der erzeugten thematischen Strukturen hängt wesentlich von der korrekten Interpretation und Zuordnung der Themen ab. Liegen die Zuordnungen vor, so müssen die daraus erstellten Zeitreihen zu einer existierenden Thematisierung passen. Die Schlüsselereignisse einer Thematisierung stellen eine Referenz außerhalb der Medien dar. Die Schlüsselereignisse werden durch ihre hohe Relevanz meist in allen Medien gleichzeitig aufgenommen und führen in der Regel zu einer erhöhten Aufmerksamkeit und Berichterstattungsmenge. Die Validität der Verfahren wird geprüft, indem die Korrelation der semi-automatisch erstellten Zeitreihen mit einer durch externe Ereignisse definierten künstlichen Zeitreihe verglichen werden. Trotz unterschiedlicher Vorbedingungen kann durch die Bestimmung der Korrelation gezeigt

werden, dass fast alle Zeitreihen eine signifikante Abhängigkeit von den Schlüsselereignissen aufweisen.[7] Die extrahierten Themenzusammenhänge der diskutierten automatischen Verfahren stehen in Bezug zu realen Ereignissen einer Thematisierung. Die wenigen Ausnahmen, welche keine Korrelation mit relevanten Ereignissen aufzeigen, sind entweder zu spezifisch oder zu generell gewählte Zusammenhänge. Die Ursache ist hier in der vorangestellten Festlegung des Themenzusammenhangs zu suchen. Die Validität der Verfahren ist grundsätzlich gegeben, wird durch die gesuchte Granularität und den Analysezweck der Themen bestimmt.

5.2.4 Weiterverarbeitung, Analyse und Anwendung von Ergebnissen

Die Möglichkeiten, die durch die Extraktion der Themenstrukturen in Dokumentkollektionen entstehen, stellen das eigentliche Potential für die inhaltsanalytische Anwendung dar. Die einem Thema zugeordneten Dokumentmengen stellen nach der semi-automatischen Verarbeitung einen strukturierten Datensatz dar. Dieser kann unterschiedlich facettiert werden und die darin enthaltenen Informationen können auf unterschiedliche Art und Weise weiterverarbeitet werden. Vor dem Hintergrund möglicher Anwendungen, die innerhalb von Themenanalysen durchgeführt werden können, sollen deren Anforderung hier noch einmal geprüft werden. Mit dem Rückbezug auf inhaltsanalytische Auswertungen und Theorien, die am Anfang der Arbeit dargestellt werden, sollen Antworten gegeben werden, wie die thematisch strukturierten Datensätze bei deren Anwendung helfen können (**F5**).

Diachrone Themenanalyse

Eine wichtige Grundlage für die Beurteilung von Themen stellt die Überführung einer Dokumentmenge in eine Zeitreihe dar. Darüber lässt sich beurteilen, wie sich die Häufigkeit eines Themas über einen Zeitraum verhält. Dies lässt Schlüsse darüber zu, welche Aufmerksamkeit einem Thema in der Öffentlichkeit zukommt. Wichtige Momente einer Thematisierung lassen sich so bestimmen und Vergleiche mit anderen Themen sind möglich. In Kapitel 4 ist dargestellt, dass die Dokumentmengen, die anhand der thematischen Zuordnung bestimmt werden, durch ihren Zeitstempel in Zeitreihen überführt werden können. Gleichseitig lässt sich die Gesamtmenge der Dokumente in unterschiedliche Zeitscheiben verteilen. Damit kann ein relativer Anteil eines Themas an einer vollständigen Berichterstattungsmenge bestimmt werden.

[7] Die Evaluation ist in Abschnitt 4.3 auf S. 143 ff. und Tabelle C.2 auf S. 209 dokumentiert.

Darüber hinaus können Dokumentmengen D_k, also thematisch verwandte Dokumentmengen, genutzt werden, um die Existenz von Eigennamen oder Wörtern in den Themen nachzuweisen. Diese können als Zeitreihe dargestellt werden. Somit kann die Frage nach diachronen Analysen so beantwortet werden, dass alle Informationen, die innerhalb der Dokumente zu einem Thema annotiert sind, genutzt werden können, um Zeitreihen aus diesen Informationen abzuleiten. Dies können insbesondere Metadaten oder Muster sein. Die Anreicherung der Dokumente mit verschiedenen Informationen kann durch das Teilgebiet der Informationsextraktion, wie der der Named Entity Recognition, realisiert werden und lässt vielfältige Anwendungen zu.

Häufigkeitsverläufe und Zyklen von Themen

Die Überführung von Dokumentmengen D_k in einen Längsschnitt der Dokumenthäufigkeiten einer Thematisierung ist die Grundlage für die Beobachtung von Themenzyklen nach Kolb (2005). Anhand dieser Zeitreihen kann ein Bezug zu den Eigenschaften unterschiedlicher Phasen einer Thematisierung gezogen werden. Kolb (2005) beschreibt unterschiedliche Phasen eines Themas, die in dieser Arbeit in Abschnitt 2.1.3 auf S. 31 ff. erläutert werden. Miltner u. Waldherr (2013) erweitern das Konzept noch auf eine spezielle Art der Thematisierung und weichen von Kolbs Idealverlauf ab. Dennoch nutzen sie die Eigenschaften der Phasen. Ausschlaggebend für die Identifikation der Phasen ist demnach der Verlauf der Dokumenthäufigkeit, die teilnehmenden Akteure und die kommunizierten Inhalte. Die Inhalte divergieren beispielsweise zwischen der bloßen deskriptiven Berichterstattung und diskursiven Auseinandersetzungen bzw. politischen Lösungsvorschlägen, die in den Medien aufgenommen werden. Über eine zusätzliche Bestimmung von Eigennamen innerhalb einer Thematisierung, so wie in Abschnitt 4.4 dargestellt, lässt sich beurteilen, welchen Anteil Personen innerhalb der Berichterstattung zu einem Thema haben. Über externe Ressourcen, wie z.B. einer externen Referenz zu einzelnen Personennamen, wäre die Trennung in politische Akteure oder wissenschaftliche Akteure realisierbar, sodass die Phasenidentifikation durch diese Informationen unterstützt werden kann. Stehen solche externen Informationen nicht zur Verfügung, bleibt dennoch der Weg für eine manuelle Zuordnung identifizierter Personennamen offen. Für die Veränderung der Bedeutung der Inhalte kann das Thema in unterschiedliche Zeitabschnitte unterteilt werden. Beispielsweise bietet sich an, die verschiedenen Brüche, wie Schlüsselereignisse, für die Abtrennung der Zeitabschnitte heranzuziehen. Über Kookkurrenzanalysen lässt sich feststellen, welche Bedeutungsverschiebungen es für die Terme in der The-

matisierung gibt. Über dieses Weg kann überprüft werden, welche Phasen ein Thema durchläuft oder durchlaufen hat.

Nachrichtenfaktoren

Die Nachrichtenfaktoren sind ein wesentlicher Bestandteil der Analyse von nachrichtenorientierten Inhalten. Sie sollen die Erklärung für die Platzierung und Aufmerksamkeit liefern, die einem Ereignis in einer Zeitung oder Nachrichtensendung zukommt. Wie in Abschnitt 2.1.3 dargestellt ist, können Faktoren wie Dauer, Nähe, Status, Dynamik oder Valenz den Wert einer Nachricht beeinflussen. Je mehr Faktoren erfüllt sind, desto prominenter ist eine Nachricht. Dennoch wird festgestellt, dass Schlüsselereignisse sich diesem Zusammenhang entziehen und meist nur in einem der Faktoren eine extreme Ausprägung aufweisen (Rauchenzauner, 2008). Besitzt eine Thematisierung an verschiedenen Tagen eine häufigere Nennung, so müssen die verantwortlichen Nachrichtenfaktoren gefunden werden, um die Thematisierung besser zu verstehen und erklären zu können. Um nun Nachrichtenfaktoren bestimmen zu können, müssen unterschiedliche Informationen aus der Dokumentmenge extrahiert werden und in einen zeitlichen Zusammenhang zur Häufigkeit der Berichterstattung gebracht werden. Einerseits muss ersichtlich sein, zu welchen Zeitpunkten eine Änderung oder Verstärkung der Häufigkeit auftritt. Andererseits muss geklärt sein, ob es sich um die Auswirkungen eines Schlüsselereignisses handelt oder ob die Berichterstattung besondere Bezüge zu oben genannten Nachrichtenfaktoren hat. Die Dokumentmenge muss demnach auf einen zu analysierenden Zeitraum eingeengt werden. Ein Beispiel wäre an dieser Stelle die Erkennung von Eigennamen im Text oder die Verwendung eines Vokabulars, dass die Existenz von Schadensbeschreibungen oder Katastrophensituationen reflektiert. Für die Bestimmung von Personennennungen ist aber zusätzliches Wissen nötig, sodass Typen oder Gruppen von Personen unterschieden werden können. Denkbar wäre hier eine deduktive Unterteilung extrahierter Person in lokale oder internationale Bekanntheitsgrade. Darüber ließe sich bestimmen, wie der Bekanntheitsgrad der Personen mit der Themenaufmerksamkeit korreliert. Dieses Vorgehen kann zeigen, wie wichtig einzelne Personen in der Presse angesehen werden. Es zeigt sich noch einmal, dass die Informationsextraktion eine wesentliche Rolle in der automatisierten Inhaltsanalyse spielt, wenn nicht nur Quantitäten von Dokumenten bestimmt werden sollen. Nochmal wird hier darauf verwiesen, dass das externe Wissen teilweise durch Analysten erstellt werden muss, um verschiedene Typen von Personen oder Wortlisten zu Schadensbeschreibungen passend zur Thematisierung zur Verfügung zu stellen. Darüber hinaus kann durch

die geeignete Längsschnittbildung gezeigt werden, wie im Laufe einer Thematisierung unterschiedliche Nachrichtenfaktoren mit der Berichterstattungsmenge korrelieren, um Langzeiteffekte bei Nachrichtenwerten zu untersuchen.

Vergleichbarkeit unterschiedlicher Quellen

In Abschnitt 4.3.2 wird gezeigt, dass die Ereignisse, die in den Quellen beschrieben werden mit unterschiedlicher Verzögerung in die Berichterstattung aufgenommen werden. Es ist auffällig, dass Online-Quellen mit einer geringen Verzögerung auf Ereignisse reagieren. Klassische Printmedien korrelieren mit den Ereignissen am besten, wenn resultierende Zeitreihen um 1 - 2 Tage verschoben werden. Dies zeigt, dass Zeitreihen abhängig von den entsprechenden Quellen sind. Somit muss bei einem Vergleich verschiedener Quellen darauf geachtet werden, dass dieser Effekt vorhanden ist. Etwaige Untersuchungen sich gegenseitig beeinflussender Medien werden nur valide, wenn der Effekt der „natürlichen“ oder normalen Verzögerung in der Berichterstattung ausgeschlossen werden kann. Wenn eine gedruckte Publikation ein Thema zwei Tage nach einer Online-Publikation aufnimmt, muss dies keinen direkten Zusammenhang haben. Es ist deswegen wichtig, dass unterschiedliche Quellen vorher überprüft und ggf. getrennt werden, sodass dieser Effekt beachtet werden kann und keinen Einfluss auf das Ergebnis hat. Stellt sich heraus, dass die analysierten Quellen einen solchen Versatz haben, so sind tageweise Vergleiche nicht mehr problemlos möglich. Bei monatlichen oder jährlichen Aggregation spielt der Effekt aber keine Rolle.

5.2.5 Datenhaltung und Datenverarbeitung

In Abschnitt 2.1.3 werden unterschiedliche Vorgehensweisen gezeigt, wie Analysen durchgeführt werden können. Dabei ist eine Unterscheidung von synchronen und diachronen Analysen zu treffen. Demnach können thematische Merkmale einer Textkollektion unabhängig oder abhängig von einer zeitlichen Komponente untersucht werden. Eine weitere Unterscheidung kann zwischen retrospektiven und prospektiven Untersuchungen getroffen werden. Wie in Kapitel 3 gezeigt wird, müssen die Dokumente, auch bei einer sequenziellen Verarbeitung, in abgeschlossenen Mengen an die Verfahren weitergegeben werden. Dies kann durch eine Stapelverarbeitung realisiert werden, die Dokumentmengen auf verschiedene Art aggregieren kann. In Kapitel 4 wird vor allem dargestellt, wie eine tageweise Verarbeitung genutzt werden kann. Es wird in diesem Kapitel aber auch gezeigt, dass die Dokumentmengen ebenfalls über eine globale Verarbeitung analysiert werden können, wenn die Analysen

retrospektiv durchgeführt werden. Damit ergeben sich für die Datenhaltung folgende Konsequenzen, um verschiedene Analysestrategien abbilden zu können (**F6**).

Um synchrone und diachrone Analysen gleichermaßen durchführen zu können, müssen für alle Dokumenten Zeitstempel hinterlegt werden. Die Darstellung der Zeit muss geeignet sein, um sinnvolle Aggregation verschiedener Zeiträume zu gewährleisten. Die Datenhaltung muss diese Aggregation anhand eines Indexes oder einer Datenstruktur direkt anbieten. Für die gezeigten Experimente werden die Dokumente als tageweise Dokumentmengen repräsentiert. Für eine sinnvolle Verknüpfung der Stapel muss die Abbildung der Wörter auf Wortnummern einer Zuordnung folgen, die für alle Dokumentmengen gilt, sodass die Inhalte und Wörter einander zuordenbar sind.

Retrospektive Analysen können auf abgeschlossenen Dokumentmengen durchgeführt werden. Die Dokumente müssen einzeln repräsentiert sein. Falls eine diachrone Betrachtung möglich sein soll, muss die Datenhaltung die Repräsentation als zeitabhängige Untermenge erlauben, damit eine Verarbeitung mit den in Kapitel 3 gezeigten Verfahren möglich ist.

Prospektive Analysen werden erst möglich, wenn die Datenhaltung in der Lage ist, neue Dokumente aufzunehmen. Es muss möglich sein, dass neue Dokumente eingegeben werden und die gleichen Vorverarbeitungen erfahren, wie alle anderen Dokumente auch. Weiterhin muss es für den Prozess möglich sein, dass neue Wörter, die bisher in keinem Dokument vorkommen, automatisch neue Wortnummern bekommen. Eine thematische Zuordnung neuer Dokumente kann erst erfolgen, wenn die Eingabe eines neuen Zeitabschnitts vollständig ist, sodass ein kompletter neuer Stapel an die Weiterverarbeitung durch die Analyseverfahren weitergegeben werden kann. Beispielsweise kann die Berichterstattung eines Tages erst vollständig mit dem letzten Tag verbunden und analysiert werden, wenn die gesamte Berichterstattung des Tages vorliegt. Hier ist es nötig einen Zeitpunkt zu bestimmen, an dem die Dokumentmenge eines Tages abgeschlossen wird. Wie in Abschnitt 3.1.2 und 3.2.4 gezeigt, ist es durch geeignete Vergleichsverfahren und Schwellwerte in Kombination mit einer Stapelverarbeitung möglich, neue Themen oder unbekannte Zusammenhänge festzustellen, sodass neue Themen entdeckt werden können und ältere Themen aus der Analyse verschwinden, wenn das Thema in wachsenden Dokumentkollektionen nicht mehr existiert oder marginalisiert wird.

An den Eigenschaften der Analyseperspektiven wird deutlich, dass die höchste Flexibilität erreicht werden kann, wenn die Datenhaltung zu jedem Dokument einen Zeitstempel kennt. Dieser muss einer Menge von Dokumenten zugeordnet werden, die

einem definierten Zeitabschnitt zugehörig sind. Die Vorverarbeitung aller Dokumente muss immer gleich sein, sodass die Repräsentation der Dokumente auf gleichen Vorbedingungen beruht. Eine Abbildung der in den Dokumenten enthaltenen Worte muss für alle Dokumente gleich sein. Im Fall der prospektiven Analysen muss die Vergabe neuer Wortnummern möglich sein, sodass ältere Dokumente anhand des erweiterten Vokabulars darstellbar werden. So können alle Verfahren für die Realisierung der Analyseperspektiven eingesetzt werden. Ein Vorschlag für die technische Realisierung einer solchen Datenhaltung und Verarbeitung wird in Anhang A gezeigt. Für die prospektive Analyse muss der methodische Ablauf auf Seite 5.1.2 insofern konkretisiert werden, dass der Import neuer Daten, deren Vorverarbeitung, die Indizierung und die Bildung neuer Dokumentmengen fortlaufend möglich ist und alle Analyseergebnisse, Themenverknüpfungen und themenabhängige Dokumentmengen, abhängig vom Analyseintervall, aktualisiert werden können. Die in Anhang A vorgestellte Softwareumgebung realisiert diese Anforderung, indem die Schritte 1-5 und 7 automatisiert in der Software abgebildet sind. Die Schritte 6 und 8 werden durch grafische Oberflächen unterstützt, sodass eine schnelle Zuordnung der Themen und eine qualitative Nachbereitung der Ergebnisse der automatischen Verfahren, wie in Abschnitt 2.2.2 gezeigt, auch in prospektiven Analysen effizient erfolgen kann. Damit sind automatisierte Verfahren der Themenanalyse ebenfalls für die Medienbeobachtung oder das Clipping in einer hohen Qualität zugänglich. Eine vollständige Automatisierung wäre im prospektiven Szenario denkbar. Die Qualität und die Validität leiden aber darunter, da durch die Verfahren induzierte Ungenauigkeiten nicht explorativ erkannt und entfernt werden.

5.3 Fazit und Ausblick

Ausgehend von den Anforderungen, die an eine Themenanalyse gestellt werden, zeigt diese Arbeit, mit welchen Methoden und Automatismen des Text-Mining diesen Anforderungen nahe gekommen werden kann. Zusammenfassend sind zwei Anforderungen herauszuheben, deren jeweilige Erfüllung die andere beeinflusst. Zum einen ist eine schnelle thematische Erfassung der Themen in einer komplexen Dokumentsammlung gefordert, um deren inhaltliche Struktur abzubilden und um Themen kontrastieren zu können. Zum anderen müssen die Themen in einem ausreichenden Detailgrad abbildbar sein, sodass eine Analyse des Sinns und der Bedeutung der Themeninhalte möglich ist. Beide Ansätze haben eine methodische Verankerung in den quantitativen und qualitativen Ansätzen der Inhaltsanalyse. Die Arbeit diskutiert diese Parallelen und setzt automatische Verfahren und Algorithmen mit den Anforderungen in Beziehung.

Von Moretti (2005) werden die Ansätze jeweils unter den Begriffen „distant-“ und „close-reading“ geführt. Es wird beschrieben, dass die Übersichtlichkeit über große Datenmengen nur gegeben ist, wenn aus einer Art Vogelperspektive, das „distant-reading“, auf die Daten geblickt wird. Details gehen verloren, aber es sind dennoch Unterschiede und verschiedene Datenpopulationen auszumachen. Im Gegensatz dazu erlaubt das „close-reading“ einen detaillierten Blick auf einen fokussierten Ausschnitt der Dokumentmenge. Analog zur explorativen Suche wird in zwei Schritten erst ein Überblick gegeben, um die Suche nach Informationen möglichst fokussiert auf die relevanten Inhalte zu lenken.

Aus dieser Unterscheidung heraus sind die Lösungen, welche in dieser Arbeit vorgeschlagen werden, entwickelt. Es können Methoden aufgezeigt werden, die eine semantische und damit thematische Trennung der Daten erlauben und einen abstrahierten Überblick über große Dokumentmengen schaffen. Dies sind Verfahren wie Topic-Modelle oder clusternde Verfahren. Mit Hilfe dieser Algorithmen ist es möglich, thematisch kohärente Untermengen in Dokumentkollektion zu erzeugen und deren thematischen Gehalt für Zusammenfassungen bereitzustellen. Es kann gezeigt werden, dass die Themen trotz der distanzierten Betrachtung unterscheidbar sind und deren Häufigkeiten und Verteilungen in einer Textkollektion diachron dargestellt werden können. Diese Aufbereitung der Daten erlaubt die Analyse von thematischen Trends oder die Selektion bestimmter thematischer Aspekte aus einer Fülle von Dokumenten. Diachrone Betrachtungen thematisch kohärenter Dokumentmengen werden dadurch möglich und die temporären Häufigkeiten von Themen können analysiert werden. Für die detaillierte Interpretation und Zusammenfassung von Themen müssen weitere Darstellungen und Informationen aus den Inhalten zu den Themen erstellt werden. Es kann gezeigt werden, dass Bedeutungen, Aussagen und Kontexte über eine Kookurrenzanalyse im Themenkontext stehender Dokumente sichtbar gemacht werden können. In einer Anwendungsform, welche die Leserichtung und Wortarten beachtet, können häufig auftretende Wortfolgen oder Aussagen innerhalb einer Thematisierung statistisch erfasst werden. Die so generierten Phrasen können zur Definition von Kategorien eingesetzt werden oder mit anderen Themen, Publikationen oder theoretischen Annahmen kontrastiert werden. Zudem sind diachrone Analysen einzelner Wörter, von Wortgruppen oder von Eigennamen in einem Thema geeignet, um Themenphasen, Schlüsselbegriffe oder Nachrichtenfaktoren zu identifizieren. Die so gewonnenen Informationen können mit einem „close-reading“ thematisch relevanter Dokumente ergänzt werden, was durch die thematische Trennung der Dokumentmengen möglich ist. Über diese methodischen Perspektiven lassen sich die automatisierten Analysen

als empirische Messinstrumente im Kontext weiterer hier nicht besprochener kommunikationswissenschaftlicher Theorien einsetzen. Des Weiteren zeigt die Arbeit, dass grafische Oberflächen und Software-Frameworks für die Bearbeitung von automatisierten Themenanalysen realisierbar und praktikabel einsetzbar sind. Insofern zeigen die Ausführungen, wie die besprochenen Lösungen und Ansätze in die Praxis überführt werden können. Die Darstellung der Potentiale automatisierter Themenuntersuchungen in großen digitalen Textkollektionen in dieser Arbeit leistet einen Beitrag zur Erforschung der automatisierten Inhaltsanalyse.

Die detaillierte Darstellung der Bedeutung und Aussagen, die in den Themen gemacht wird, ist eine wichtige Grundlage, um deren Darstellung an unterschiedlichen Zeitpunkten oder in verschiedenen Publikationen unterscheidbar zu machen. Um eine detaillierte Unterscheidung und die Extraktion von Aussagen zu unterstützen, wird die Kookkurrenzanalyse als Hilfsmittel vorgeschlagen. Die komplette Abstraktion statistisch signifikanter Textaussagen für die Produktion von Makropropositionen von Mackeldey (1987) oder zur Kategorienbildung (Früh, 2001) ist allerdings im Rahmen dieser Arbeit noch nicht ausreichend untersucht. Hier kann ein Potential für die künftige Forschung identifiziert werden. Um automatisiert Propositionen aus Texten extrahieren zu können, müssen automatische Satzanalysen in der Form durchgeführt werden, dass Funktionen im Satz bestimmt werden können. Aus der Theorie ist bekannt, dass elementare Bausteine einer Proposition aus Substantiv, Verb und Adjektiv-Adverbial Bestandteilen eines Satzes konstruiert werden (Hausser, 2000, vgl. S. 66). An diese Funktionen muss über Syntaxbäume herangegangen werden. Mit einer statistischen Analyse, die diese Satzfunktionen beachtet, ist es möglich, statistisch signifikante Beziehungen von Wörtern bestimmter Funktionen zu berechnen. Bestehen die Funktionen eines Satzes aus Phrasen mehrerer Wörter müssen diese Phrasen identifiziert werden. Die Kookkurrenzanalyse muss von der Wortebene auf die Ebene der elementaren Bausteine einer Proposition überführt werden. Ein Baustein oder Bestandteil einer Proposition kann mit mehreren anderen Bausteinen in propositionaler Beziehung stehen, was wiederum zu einem Netzwerk von verknüpften Aussagen führt. Es kann zu einem Objekt mehrfach etwas ausgesagt werden. Werden alle Bausteine nun als Sammlung aufgefasst, so kann der Versuch gemacht werden, in Beziehung stehende Bausteine zu abstrahieren. Eine Waffe kann im Text durch Nennungen wie Pistole, Messer oder Gewehr genannt werden. Soll aus zwei Propositionen, die jeweils ähnliche Aussagen enthalten, abstrahiert werden, muss mit externem Wissen dafür gesorgt werden, dass einzelne Ausprägungen der Bestandteile einer Proposition so zusammengefasst werden können. Mit einer geeigneten Wissens-

basis ist dieser Schritt durchaus denkbar. Ressourcen wie WordNET (Fellbaum, 1998) eignen sich als Grundlage für die Abstraktion ähnlicher Propositionen. Dies wird einen entscheidenden Beitrag für die Erfassung kategorialer Ausprägungen in großen Dokumentkollektionen leisten.

Mit dieser Idee für eine automatisierte Bedeutungsanalyse soll abschließend gezeigt werden, dass neben der reinen Thementrennung und der detaillierten Analyse der Themen durch automatisierte Methoden, weitere Perspektiven für die Weiterentwicklung automatisierter kommunikations- und sozialwissenschaftlicher Themen- und Inhaltsanalysen existieren.

Anhang A

Software und Verarbeitungsabläufe

Die Software, die für die semi-automatische Themenanalyse von digitalen Textquellen benötigt wird, muss 4 Teilmodule für

1. die **Datenbank**,
2. die **Vorverarbeitung**,
3. die Berechnung der **Themen** und
4. die nachträgliche **Auswertung** der Themen enthalten.

Eine Datenbank wird vor allem benötigt, um die Rohtexte und die verarbeiteten Korpora vorzuhalten. Stehen die zu analysierenden Inhalte nicht als digitale Dokumentkollektion zur Verfügung, so muss das Material eventuell aus Online-Inhalten erstellt werden. Dabei kann oft auf Schnittstellen zugegriffen werden, die den Zugriff auf die Inhalte gewährleisten.[1] Ist der Zugriff über eine Schnittstelle nicht gegeben, so müssen die Inhalte von Internetseiten abgegriffen werden. Dies erfordert allerdings eine Säuberung des HTML-Markups, sodass ein weiterer Schritt für die Verarbeitung der Daten nötig ist. Dieser ist nicht unproblematisch, da es schwer ist, relevanten Inhalt von Zusatzinformationen oder Verweisen auf andere Artikel zu trennen.

Zunächst müssen die gewonnenen Inhalte für die Themenanalyse vorbereitet werden. Dazu müssen die Texte in ein Korpusformat überführt werden. Im Fall der

[1] Ein Beispiel für eine API, die den Zugriff auf Online-Nachrichten gewährt, ist die Guardian Open Plattform (Guardian).

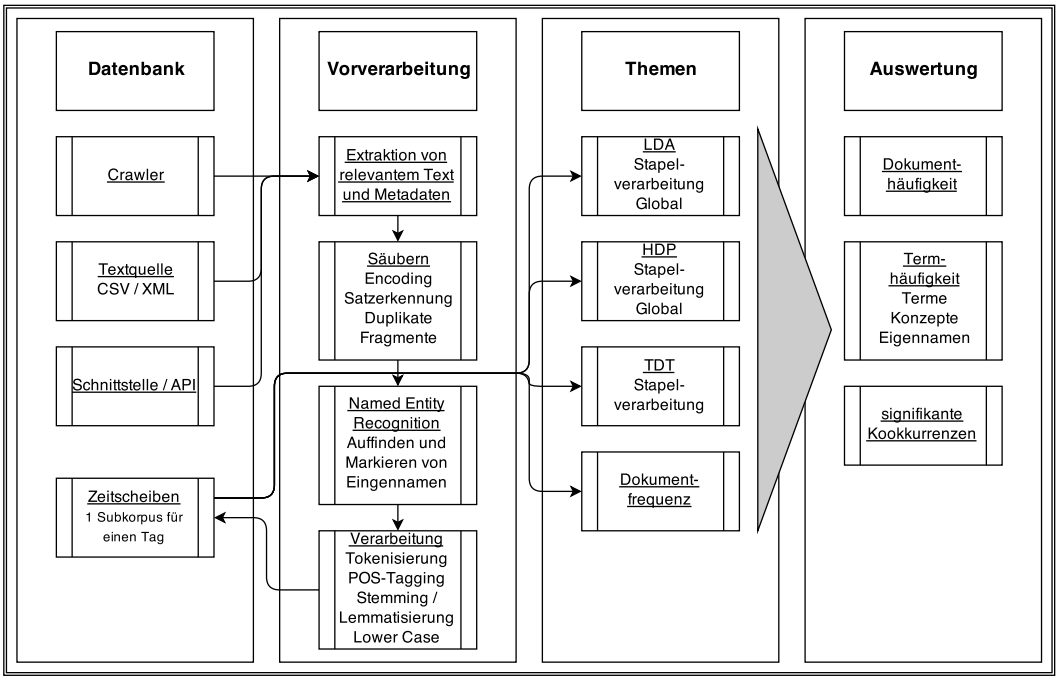

Abbildung A.1: Schematische Darstellung der Softwarekomponenten und der Verarbeitungskette.

Themenanalysen müssen in den Quelldaten relevante Texte und Metadaten extrahiert werden. Informationen wie Veröffentlichungsdatum, Autoren, Quellenname und die Dokumentüberschrift sind für eine spätere Auswertung der Artikel essentiell. Ohne die Zeitstempel der Dokumente kann keine diachrone Analyse stattfinden. Weiterhin ist die einheitliche Verwendung eines Zeichensatzes, wie z.B. UTF-8, wichtig, um die korrekte und einheitliche Darstellung von Sonderzeichen zu gewährleisten. Der erste Schritt, um einen Korpus zu erstellen, ist die Markierung vollständiger Sätze in den Texten. So ist es möglich die Texte Satz für Satz auszuwerten. Über diesen Weg können doppelte Sätze (Duplikate) eliminiert werden, die eine korrekte Verarbeitung in den Verfahren stören würden. Darüber hinaus kann eine strikte Filterung unvollständiger Sätze vorgenommen werden, um eventuelle Fragmente in den Texten unbeachtet zu lassen. Die so vorbereiteten Texte können Satz für Satz tokenisiert werden und den Token kann ein POS-Tag zugeteilt werden. Für eine Reduktion der Wortformen kommt zusätzlich eine Lemmatisierung und eine Transformation aller Wortformen in Kleinschreibung zum Einsatz. Für die Vorverarbeitung wurden teilweise vorbereitete Bibliotheken und Ressourcen verwendet. Für die Verkettung der Verfahren und den Zugriff auf die Datenstruktur wurden, nach dem Vorbild des schematischen Aufbaus in Abbildung A.1, Prozessketten in Java implementiert. Die so verknüpften Programme, die teilweise mit existierenden Bibliotheken implementiert wurden, können als Framework für die Themenextraktion verstanden werden. Im Folgenden werden die

dafür verwendeten Ressourcen und Programmbibliotheken genannt, die nicht selbst entwickelt wurden.

- Satzerkennung, Tokenisierung, POS Tagging - openNLP (openNLP)
- Named Entity Recognition - StanfordNLP (Finkel u. a., 2005)
- Lemmatisierung - Liste von Grundformen aus dem Projekt Deutscher Wortschatz (wortschatz 1)
- Zusammenstellung des Guardian Corpus - Guardian Open Plattform (Guardian)
- Zusammenstellung der Korpora SZ und TAZ - Digitale Archive der Verlage (nicht öffentlich)

Während der Vorverarbeitung werden alle Dokumente einer Publikation, die am gleichen Tag erstellt wurden, in ein Sub-Korpus überführt. Damit entsteht eine Datenbank, die eine sequenzielle Abfolge von Korpora enthält. Alle Zeitscheiben gemeinsam definieren den Gesamtkorpus einer jeweiligen Publikation wie beispielsweise einer Tageszeitung. Die Zuordnung der Dokumente zu Zeitscheiben wird benötigt, um die sequenzielle Verarbeitung der Dokumente zu simulieren. Mit dieser Funktionalität ist gewährleistet, dass die Verarbeitung in Echtzeit, das heißt mit einem kontinuierlichen täglichen Datenstrom, funktionieren würde. In den Zeitscheiben können jeweils Statistiken berechnet werden, sodass beispielsweise Worthäufigkeiten für die tagesaktuellen Korpora angegeben werden können. Für das Format des Korpus ist die geeignete Ablage der Dokumente wichtig, welche die Dokumentstruktur abzubilden vermag. Die Trennung der Dokumente in Absätze oder Sätze ist besonders wichtig und wird über die zeilenweise Speicherung als Tupel in einer Datenbanktabelle realisiert. Zusammengehörige Sätze oder Absätze werden durch einen Dokumentbezeichner (D_ID) gekennzeichnet. Darüber können dokumentzugehörige Metadaten verknüpft werden, die in einer anderen Tabelle vorgehalten werden. Das zeilenweise Format wäre über das Tupel $\{ID, TEXT, D\_ID\}$ abgebildet und eine weitere Verarbeitung unter Berücksichtigung von Sätzen oder Absätzen ist möglich.

Die zeitscheibenbasierte Verarbeitung wird einerseits in Form einer Stapelverarbeitung jeder Zeitscheibe und andererseits als globale Verarbeitung aller Textdaten realisiert. Für eine sequenzielle und prospektive Analyse ist allerdings nur die Stapelverarbeitung praktikabel. Die Verfahren LDA, HDP und TDT wurden den Veröffentlichungen folgend in Java implementiert. Bei den Topic-Modellen wurde jeweils ein

Gibbs Sampler als Inferenzverfahren gewählt (Heinrich, 2005; Griffiths u. Steyvers, 2004; Teh u. Jordan, 2010). Die Auswertung der Dokumentfrequenz einer Menge von Schlüsselbegriffen in den Zeitscheiben wurde innerhalb des hier vorgeschlagenen Prozessablaufs in Java implementiert. Für die Verarbeitung der Korpora mit den Verfahren kann im Gesamtaufbau des Systems eine Wortliste mit zu ignorierenden Wörtern angegeben werden, sodass beispielsweise Stopwörter aus der Verarbeitung ausgeschlossene werden können. Zusätzlich ist es wichtig, einen Schwellwert für die Mindesthäufikeit eines Terms in einer Zeitscheibe anzugeben. Dies verhindert die Einbeziehung statistisch irrelevanter Wortformen und beschleunigt die Verfahren. Die Ergebnisse werden in einer Datenstruktur gespeichert, die eine Zuordnung der Dokumente zu einem Thema herstellt. So ist es über eine Tabelle einfach möglich, jedem Bezeichner eines Themas (K) eine Dokumentmenge zuzuordnen. Die Ergebnisse für eine feste Zuordnung von Dokumenten zu Themen sind über das Tupel $\{K, D_k\}$ abgebildet, wobei D_k eine Liste aller Dokumentbezeichner D_ID darstellt, die dem Thema K zugeordnet sind. Die Abbildung der Topic-Modelle bedarf einer komplexeren Datenstruktur, um die Verteilungen $p(\mathbf{w}|z)$ und $p(\mathbf{z}|d)$ abzubilden. Die Verteilungen in den Dokumenten lassen sich für ein Modell durch das Tupel $\{D\_ID, W_D, K_D\}$ abbilden. Dabei ist W_D die sortierte Liste aller Wortformen, die mit ihrem Bezeichner, also eine Referenz auf eine Wortliste des Korpus, angegeben werden. Die Menge K_D gibt das für jedes Wort zugewiesene Thema K an. Aus diesen Angaben lässt sich die Verteilung $p(\mathbf{z}|d)$ für jedes Dokument wieder herstellen. Die Repräsentation der Verteilung $p(\mathbf{w}|z)$ in einem Topic-Modell, muss eine Wahrscheinlichkeit für jede Wortform innerhalb eines Themas enthalten. Das Tupel $\{K, W, P\}$ repräsentiert diese Datenstruktur. Mit der Liste W aller Wortformen, die auf eine Wahrscheinlichkeit P im Thema K abgebildet werden, kann dies realisiert werden. Im Fall der Verfahren LDA und HDP reichen die Datenstrukturen aus, um ein Modell ausreichend zu beschreiben. Das Verfahren TDT muss in einer Datenstruktur repräsentiert werden, die alle Dokumente als Liste von 1000 Wortformen und dem assoziierten Termgewicht repräsentiert. Es ist nützlich, die Wortformen mit ihrem Bezeichner zu hinterlegen. Das Tupel $\{D\_ID, W_D, TW_D\}$ beschreibt alle Dokumente als limitierte Liste von Wortformen W_D mit der zugehörigen Liste von Termgewichten TW. Die Größe der Menge W_D wird auf maximal 1000 Elemente für ein Dokument beschränkt.

Die Auswertung berechneter Themenstrukturen erfolgt durch Module, die auf die Datenstruktur der automatischen Verfahren zugreifen können. Dies basiert einerseits auf der Dokumentmenge D_k eines Themas und andererseits, im Fall der Topic-Modelle, auf der Wahrscheinlichkeit, die jedem Thema in den Dokumenten zugeordnet ist. Die

Zusammenfassungen der Themen, die als Termvektoren angegeben werden, können durch die Repräsentation der Modelle und Ergebnisse als geeignete Tupel aus den Daten erzeugt werden. Die Zusammenfassung der Topic-Modelle lässt sich trivial durch Sortierung der Wahrscheinlichkeiten und Wortformen aus der Datenstruktur $\{K, W, P\}$ herstellen. Die Zusammenfassungen für das Verfahren TDT kann über die Berechnung eines Durchschnittsvektors für alle Dokumente eines Themas hergestellt werden. Durch die Repräsentation der Dokumente als Liste von 1000 Wortformen mit zugehörigem Termgewicht, müssen diese Daten entsprechend aus der Datenstruktur verarbeitet werden. Für alle weiteren Berechnungen, wie die Konkurrenzanalysen oder die Betrachtung themenabhängiger Worthäufigkeiten, ist jeweils die Dokumentmenge D_k relevant. So können Berechnungen von Worthäufigkeiten, Häufigkeiten von Eigennamen oder Konkurrenzen nachträglich auf den Themenstrukturen durchgeführt werden.

Durch die geeignete Repräsentation der Korpora und der darin enthaltenen Themen können Visualisierungen direkt auf den Ergebnissen einer Themenanalyse aufgebaut werden. Die Interpretation der gespeicherten Tupel als gewichtete Liste von Wörtern lässt sich so in visuelle Eigenschaften überführen. Die in der Arbeit genutzten Visualisierungen basieren auf unterschiedlichen Implementierungen, die auf der Grundlage der vorgestellten Datenstruktur entwickelt wurden. Die Visualisierungen wurden eigens implementiert und in verschiedenen Beiträgen, die die Visualisierung und interaktive Nutzung von Themenanalysen demonstrieren, vorgestellt (Niekler u. a., 2012, 2014b). Niekler u. a. (2014b) haben die dargestellt Prozesskette für Inhalts- und Themenanalysen in einer Softwarearchitektur implementiert.

Anhang B

Beispielhafte Textdateien für die explorative Themenanalyse

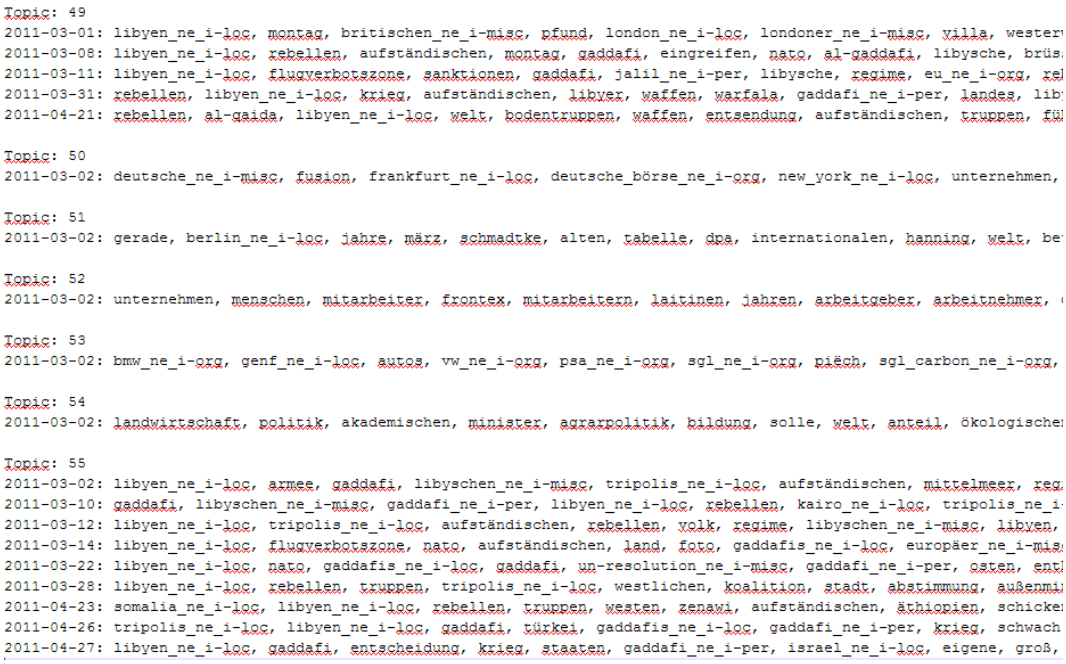

Abbildung B.1: Textdarstellung verketteter Themen. Die Themen wurden innerhalb von Tageszeitscheiben von je einem LDA Modell erstellt. Verkettete Themen werden zeilenweise innerhalb einer eindeutigen Bezeichnung für das Thema zusammengefasst.

Abbildung B.2: Darstellung einer globalen Berechnung eines Topic Modells. Jede Zeile repräsentiert einen Wortvektor eines globalen LDA-Modells als thematische Zusammenfassung. Zeilen, die bestimmte Worte enthalten, können zur leichteren Identifikation markiert werden.

Abbildung B.3: Darstellung einer TDT-Berechnung. Jede Zeile repräsentiert einen Wortvektor als thematische Zusammenfassung und die zugeordneten Dokumente. Auch hier können Zeilen, die bestimmte Worte enthalten zur leichteren Identifikation markiert werden.

Anhang C

Tabellen

	taz	sz	guardian	taz_base	sz_base
lda.ClusterTimeSeries/lib_main/global_low	0.030	0.031	0.018	0.029	0.034
lda.ClusterTimeSeries/lib_main/global_hi	0.025	0.010	0.014	0.027	0.014
lda.ClusterTimeSeries/lib_main/05	0.030	0.024	0.036	0.029	0.025
lda.ClusterTimeSeries/lib_all/global_low	0.039	0.039	0.046	0.039	0.045
lda.ClusterTimeSeries/lib_all/global_hi	0.030	0.021	0.033	0.032	0.029
lda.ClusterTimeSeries/lib_all/05	0.037	0.035	0.044	0.035	0.036
lda.ClusterTimeSeries/fuk_all/global_low	0.087	0.054	0.037	0.081	0.069
lda.ClusterTimeSeries/fuk_all/global_hi	0.049	0.043	0.025	0.050	0.050
lda.ClusterTimeSeries/fuk_all/05	0.075	0.064	0.022	0.077	0.064
lda.ClusterTimeSeries/fuk_all/06	0.067	0.056	0.019	0.075	0.063
lda.ClusterTimeSeries/fuk_main/global_low	0.042	0.041	0.019	0.040	0.036
lda.ClusterTimeSeries/fuk_main/global_hi	0.011	0.013	0.007	0.025	0.025
lda.ClusterTimeSeries/fuk_main/05	0.032	0.042	0.017	0.032	0.034
lda.ClusterTimeSeries/fuk_main/06	0.026	0.028	0.016	0.031	0.032
hdp_lda.ClusterTimeSeries/lib_main/global_low	0.053	0.061	0.022	0.097	0.092
hdp_lda.ClusterTimeSeries/lib_main/global_hi	0.025	0.044	0.015	0.035	0.028
hdp_lda.ClusterTimeSeries/lib_main/05	0.024	0.026	0.045	0.027	0.027
hdp_lda.ClusterTimeSeries/lib_all/global_low	0.065	0.061	0.052	0.097	0.111
hdp_lda.ClusterTimeSeries/lib_all/global_hi	0.034	0.072	0.033	0.044	0.041
hdp_lda.ClusterTimeSeries/lib_all/05	0.030	0.031	0.052	0.038	0.036
hdp_lda.ClusterTimeSeries/fuk_all/global_low	0.107	0.077	0.038	0.093	0.071
hdp_lda.ClusterTimeSeries/fuk_all/global_hi	0.081	0.054	0.025	0.072	0.063
hdp_lda.ClusterTimeSeries/fuk_all/05	0.068	0.058	0.016	0.068	0.060
hdp_lda.ClusterTimeSeries/fuk_all/06	0.074	0.059	0.016	0.075	0.054
hdp_lda.ClusterTimeSeries/fuk_main/global_low	0.041	0.077	0.020	0.051	0.042
hdp_lda.ClusterTimeSeries/fuk_main/global_hi	0.017	0.032	0.015	0.034	0.033
hdp_lda.ClusterTimeSeries/fuk_main/05	0.028	0.033	0.015	0.035	0.036
hdp_lda.ClusterTimeSeries/fuk_main/06	0.028	0.020	0.014	0.030	0.033
lda.ProbTimeSeries/lib_main/global_low	0.026	0.022	0.022	0.025	0.023
lda.ProbTimeSeries/lib_main/global_hi	0.022	0.009	0.019	0.027	0.009
lda.ProbTimeSeries/lib_main/05	0.030	0.019	0.045	0.028	0.018
lda.ProbTimeSeries/lib_all/global_low	0.033	0.028	0.054	0.034	0.031
lda.ProbTimeSeries/lib_all/global_hi	0.027	0.017	0.044	0.031	0.020
lda.ProbTimeSeries/lib_all/05	0.038	0.027	0.055	0.034	0.025
lda.ProbTimeSeries/fuk_all/global_low	0.077	0.043	0.032	0.071	0.053
lda.ProbTimeSeries/fuk_all/global_hi	0.047	0.039	0.026	0.049	0.041
lda.ProbTimeSeries/fuk_all/05	0.075	0.055	0.022	0.072	0.055
lda.ProbTimeSeries/fuk_all/06	0.068	0.051	0.021	0.070	0.055
lda.ProbTimeSeries/fuk_main/global_low	0.036	0.031	0.020	0.036	0.026
lda.ProbTimeSeries/fuk_main/global_hi	0.009	0.012	0.006	0.026	0.020
lda.ProbTimeSeries/fuk_main/05	0.034	0.035	0.019	0.029	0.030
lda.ProbTimeSeries/fuk_main/06	0.027	0.025	0.018	0.030	0.028
hdp_lda.ProbTimeSeries/lib_main/global_low	0.049	0.049	0.027	0.083	0.064
hdp_lda.ProbTimeSeries/lib_main/global_hi	0.021	0.029	0.020	0.031	0.019
hdp_lda.ProbTimeSeries/lib_main/05	0.028	0.022	0.059	0.028	0.021
hdp_lda.ProbTimeSeries/lib_all/global_low	0.057	0.049	0.067	0.083	0.079
hdp_lda.ProbTimeSeries/lib_all/global_hi	0.030	0.056	0.041	0.038	0.032
hdp_lda.ProbTimeSeries/lib_all/05	0.034	0.027	0.067	0.038	0.028
hdp_lda.ProbTimeSeries/fuk_all/global_low	0.081	0.059	0.033	0.079	0.054
hdp_lda.ProbTimeSeries/fuk_all/global_hi	0.063	0.043	0.024	0.067	0.049
hdp_lda.ProbTimeSeries/fuk_all/05	0.074	0.056	0.020	0.067	0.052
hdp_lda.ProbTimeSeries/fuk_all/06	0.081	0.056	0.021	0.075	0.049
hdp_lda.ProbTimeSeries/fuk_main/global_low	0.040	0.059	0.022	0.045	0.032
hdp_lda.ProbTimeSeries/fuk_main/global_hi	0.015	0.025	0.017	0.032	0.024
hdp_lda.ProbTimeSeries/fuk_main/05	0.033	0.030	0.019	0.035	0.030
hdp_lda.ProbTimeSeries/fuk_main/06	0.033	0.019	0.019	0.031	0.028
tdt.ClusterTimeSeries/lib_main/03	0.034	0.036	0.047	0.033	0.035
tdt.ClusterTimeSeries/lib_all/03	0.045	0.036	0.051	0.048	0.050
tdt.ClusterTimeSeries/fuk_all/03	0.068	0.053	0.025	0.080	0.062
tdt.ClusterTimeSeries/fuk_main/03	0.053	0.039	0.025	0.076	0.055

Tabelle C.1: Anteil S_k^D und P_k^D der Themen LIB und FUK ermittelt durch verschiedene Verfahren in unterschiedlichen Korpora.

	taz	sz	guardian	taz_base	sz_base
lda.ClusterTimeSeries/lib_main/global_low	0.522	0.540	0.319	0.497	0.513
lda.ClusterTimeSeries/lib_main/global_hi	0.534	0.164	0.240	0.512	0.341
lda.ClusterTimeSeries/lib_main/05	0.531	0.549	0.596	0.494	0.676
lda.ClusterTimeSeries/lib_all/global_low	0.462	0.498	0.585	0.355	0.477
lda.ClusterTimeSeries/lib_all/global_hi	0.447	0.456	0.547	0.415	0.496
lda.ClusterTimeSeries/lib_all/05	0.490	0.541	0.545	0.494	0.525
lda.ClusterTimeSeries/fuk_all/global_low	0.447	0.556	0.596	0.461	0.542
lda.ClusterTimeSeries/fuk_all/global_hi	0.448	0.527	0.709	0.506	0.539
lda.ClusterTimeSeries/fuk_all/05	0.427	0.514	0.662	0.496	0.616
lda.ClusterTimeSeries/fuk_all/06	0.489	0.542	0.648	0.433	0.524
lda.ClusterTimeSeries/fuk_main/global_low	0.543	0.540	0.701	0.459	0.505
lda.ClusterTimeSeries/fuk_main/global_hi	0.596	0.546	0.638	0.519	0.503
lda.ClusterTimeSeries/fuk_main/05	0.556	0.515	0.635	0.545	0.515
lda.ClusterTimeSeries/fuk_main/06	0.536	0.536	0.663	0.551	0.485
hdp_lda.ClusterTimeSeries/lib_main/global_low	0.415	0.497	0.177	0.200	0.349
hdp_lda.ClusterTimeSeries/lib_main/global_hi	0.591	0.326	0.385	0.487	0.484
hdp_lda.ClusterTimeSeries/lib_main/05	0.404	0.668	0.457	0.469	0.551
hdp_lda.ClusterTimeSeries/lib_all/global_low	0.388	0.497	0.512	0.200	0.288
hdp_lda.ClusterTimeSeries/lib_all/global_hi	0.533	0.415	0.588	0.381	0.469
hdp_lda.ClusterTimeSeries/lib_all/05	0.351	0.681	0.382	0.388	0.551
hdp_lda.ClusterTimeSeries/fuk_all/global_low	0.432	0.522	0.622	0.465	0.553
hdp_lda.ClusterTimeSeries/fuk_all/global_hi	0.459	0.538	0.658	0.465	0.550
hdp_lda.ClusterTimeSeries/fuk_all/05	0.505	0.558	0.599	0.432	0.521
hdp_lda.ClusterTimeSeries/fuk_all/06	0.441	0.623	0.583	0.382	0.549
hdp_lda.ClusterTimeSeries/fuk_main/global_low	0.514	0.522	0.680	0.507	0.509
hdp_lda.ClusterTimeSeries/fuk_main/global_hi	0.588	0.545	0.709	0.501	0.533
hdp_lda.ClusterTimeSeries/fuk_main/05	0.590	0.519	0.616	0.527	0.530
hdp_lda.ClusterTimeSeries/fuk_main/06	0.592	0.406	0.636	0.552	0.524
lda.ProbTimeSeries/lib_main/global_low	0.520	0.553	0.476	0.506	0.512
lda.ProbTimeSeries/lib_main/global_hi	0.523	0.314	0.281	0.463	0.460
lda.ProbTimeSeries/lib_main/05	0.537	0.541	0.637	0.461	0.659
lda.ProbTimeSeries/lib_all/global_low	0.465	0.550	0.634	0.424	0.545
lda.ProbTimeSeries/lib_all/global_hi	0.450	0.479	0.616	0.417	0.594
lda.ProbTimeSeries/lib_all/05	0.536	0.589	0.606	0.505	0.607
lda.ProbTimeSeries/fuk_all/global_low	0.488	0.578	0.721	0.482	0.560
lda.ProbTimeSeries/fuk_all/global_hi	0.455	0.561	0.746	0.541	0.573
lda.ProbTimeSeries/fuk_all/05	0.481	0.551	0.727	0.481	0.615
lda.ProbTimeSeries/fuk_all/06	0.497	0.550	0.726	0.480	0.544
lda.ProbTimeSeries/fuk_main/global_low	0.575	0.556	0.738	0.538	0.534
lda.ProbTimeSeries/fuk_main/global_hi	0.559	0.552	0.648	0.556	0.549
lda.ProbTimeSeries/fuk_main/05	0.570	0.566	0.716	0.559	0.532
lda.ProbTimeSeries/fuk_main/06	0.558	0.546	0.731	0.568	0.506
hdp_lda.ProbTimeSeries/lib_main/global_low	0.479	0.542	0.298	0.278	0.367
hdp_lda.ProbTimeSeries/lib_main/global_hi	0.549	0.388	0.470	0.464	0.508
hdp_lda.ProbTimeSeries/lib_main/05	0.399	0.641	0.543	0.509	0.567
hdp_lda.ProbTimeSeries/lib_all/global_low	0.432	0.542	0.582	0.278	0.345
hdp_lda.ProbTimeSeries/lib_all/global_hi	0.485	0.510	0.629	0.421	0.482
hdp_lda.ProbTimeSeries/lib_all/05	0.347	0.683	0.503	0.418	0.537
hdp_lda.ProbTimeSeries/fuk_all/global_low	0.507	0.545	0.717	0.513	0.564
hdp_lda.ProbTimeSeries/fuk_all/global_hi	0.490	0.558	0.731	0.477	0.550
hdp_lda.ProbTimeSeries/fuk_all/05	0.552	0.579	0.683	0.493	0.545
hdp_lda.ProbTimeSeries/fuk_all/06	0.473	0.626	0.677	0.435	0.566
hdp_lda.ProbTimeSeries/fuk_main/global_low	0.565	0.545	0.736	0.558	0.517
hdp_lda.ProbTimeSeries/fuk_main/global_hi	0.570	0.554	0.743	0.560	0.551
hdp_lda.ProbTimeSeries/fuk_main/05	0.586	0.539	0.691	0.567	0.551
hdp_lda.ProbTimeSeries/fuk_main/06	0.604	0.447	0.699	0.593	0.552
tdt.ClusterTimeSeries/lib_main/03	0.287	0.355	0.294	0.188	0.352
tdt.ClusterTimeSeries/lib_all/03	0.373	0.355	0.365	0.258	0.237
tdt.ClusterTimeSeries/fuk_all/03	0.398	0.427	0.660	0.433	0.460
tdt.ClusterTimeSeries/fuk_main/03	0.392	0.503	0.660	0.440	0.473

Tabelle C.2: Kreuzkorrelation der Themenzeitreihen aus unterschiedlichen Verfahren, Parametrisierungen und Korpora. Aus der Kreuzkorrelation wurde jeweils das Lag identifiziert, welches die höchste Korrelation hervorruft. Der Korrelationswert wurde dann für jedes Verfarhen und jeden Korpus eingetragen.

Anhang D

Algorithmen

Algorithm 1 Die sequenzielle Berechnung von Story-Clustern mit Formel 3.1.

$D \leftarrow$ sortierte Dokumente
Initialisiere TF und DF mit 0.
while nicht verarbeiteter Batch existiert **do**
  Selektiere ältesten nicht verarbeiteten Batch t.
  Definiere eine leere Menge bereits verarbeiteter Dokumente p.
  Initialisiere/Aktualisiere TF und DF anhand der Dokumente im Batch t.
  for all Dokument A im Batch t **do**
    Berechne Termgewichte nach Formel 3.1 für A.
    Erstelle einen Termvektor für die Named-Entity-Terme und einen Vektor für die restlichen Terme (Themen- oder Story-Terme).
    $S \leftarrow 0$
    for all Dokument B in p **do**
      Vergleiche A und B mit Formel 3.2 und ermittle
      $S \leftarrow \text{maxsim}B$.
    if S $>$ Schwellwert **then**
      Setze Cluster von A auf Cluster von B mit der höchsten Ähnlichkeit.
    else
      Initialisiere einen neuen Cluster für A.
    Füge A der Menge p hinzu

Algorithm 2 Sequenzielle Berechnung von Themenähnlichkeiten innerhalb verschiedener Zeitscheiben (Batch) mit Topic Modellen.

$h \leftarrow$ Anzahl der zu beachtenden vorhergehende Batches
$Kandidat \leftarrow \emptyset$
while nicht verarbeiteter Batch existiert **do**
  Selektiere ältesten nicht verarbeiteten Batch t
  $History \leftarrow$ Selektiere aufsteigend sortierte Batches, beginnend mit Batch $t - h$.
  $Topics_t \leftarrow$ Selektiere alle Topics aus Batch t.
  for all Batch $b \in History$ **do**
    $Topics_h \leftarrow$ Selektiere alle Topics aus Batch b.
    for all $A \in Topics_t$ **do**
      for all $B \in Topics_h$ **do**
        $sim(A, B) =$ Formel 3.8
        if $Kandidat(A) \neq \emptyset$ **then**
          if $sim(A, B) > sim(A, Kandidat(A))$ **then**
            if $B = Kandidat(A'), A' \in Topics_t$ **then**
              if $sim(A, B) > sim(A', B)$ **then**
                $Kandidat(A) \leftarrow B$
                Lösche den Eintrag$Kandidat(A')$.
            else
              $Kandidat(A) \leftarrow B$
        else
          if $sim(A, B) >$ Schwellwert **then**
            if $B = Kandidat(A'), A' \in Topics_t$ **then**
              if $sim(A, B) > sim(A', B)$ **then**
                $Kandidat(A) \leftarrow B$
                Lösche den Eintrag$Kandidat(A')$.
            else
              $Kandidat(A) \leftarrow B$
    for all $A \in Topics_t$ **do**
      if $Kandidat(A)$ **then**
        Füge A der Menge von Themen zu, der $Kandidat(A)$ zugeordnet ist.
      else
        Erzeuge eine neue leere Menge und füge A in diese Menge ein.

Algorithm 3 Vergleichsverfahren für die Überprüfung der Reliabilität von Topic Modellen

$TM \leftarrow$ 10 Topic Modelle
$comp \leftarrow 10 \times 10$ Matrix
for all Topic Modell $i \in TM$ **do**
  $maxsims \leftarrow$ leere Liste
  for all Topic Modell $j \in TM$ **do**
    for all Topic $ti \in i$ **do**
      for all Topic $tj \in j$ **do**
        Vergleiche tiundtjmit Formel 3.2
      $max \leftarrow$ maxsim(ti)
      Füge max zu $maxsims$ hinzu
    $count \leftarrow$ Zähle Elemente in $maxsims$ für die < 0.5 gilt
    $comp(i, j) = count$
$u \leftarrow \overline{lower.tri(comp)}$
$error \leftarrow u$ / 50

Literaturverzeichnis

[Ágel 2000] Ágel, Vilmos: *Valenztheorie.* Tuübingen: Narr, 2000

[Agricola 1976] Agricola, Erhard: Vom Text zum Thema. In: *Probleme der Textgrammatik I.* Berlin: Akademie-Verlag, 1976, S. 13–27

[Aitchison u. Shen 1980] Aitchison, J.; Shen, S. M.: Logistic-Normal Distributions: Some Properties and Uses. In: *Biometrika* 67 (1980), Nr. 2, S. 261–272

[Allan 2002] Allan, James: *Topic detection and tracking: Event-based information organization.* Boston: Kluwer Acad. Publ., 2002

[Allan u. a. 1998] Allan, James; Carbonell, Jaime; Doddington, George; Yamron, Jonathan ; Yang: Topic Detection and Tracking Pilot Study: Final Report. In: *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Morgan Kaufmann, 1998, S. 194–218

[Allan u. a. 2005] Allan, James; Harding, Stephen; Fisher, David; Bolivar, Alvaro; Guzman-Lara, Sergio ; Amstutz, Peter: Taking Topic Detection From Evaluation to Practice. In: *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society, 2005 (HICSS '05)

[Alpaydin 2008] Alpaydin, Ethem: *Maschinelles Lernen.* München: Oldenbourg, 2008

[Asuncion u. a. 2009] Asuncion, Arthur; Welling, Max; Smyth, Padhraic ; Teh, Yee W.: On Smoothing and Inference for Topic Models. In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, 2009 (UAI '09), S. 27–34

[Baeza-Yates u. Ribeiro-Neto 2011] Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier: *Modern information retrieval.* New York: Addison-Wesley, 2011

[Bayer 1980] Bayer, Josef: Diskursthemen. In: Weigand, E. (Hrsg.); Tschauder, G. (Hrsg.): *Perspektive: textintern: Akten des 14. Linguistischen Kolloquiums Bochum 1979.* Tübingen: M. Niemeyer, 1980, S. 216–224

[Berelson 1984] Berelson, Bernard: *Content analysis in communication research.* New York: Hafner Press, 1984

[Bergenholtz u. Schaeder 1977] Bergenholtz, Henning; Schaeder, Burkhard: *Die Wortarten des Deutschen: Versuch e. syntakt. orientierten Klassifikation.* Stuttgart: Klett, 1977

[Biemann 2012] Biemann, Chris: *Structure discovery in natural language.* Berlin: Springer, 2012

[Bilandzic u. a. 2001] Bilandzic, Helena; Koschel, Friederike ; Scheufele, Bertram: Theoretisch-heuristische Segmentierung im Prozeß der empiriegeleiteten Kategorienbildung. In: *Inhaltsanalyse: Perspektiven, Probleme, Potentiale.* Köln: Halem, 2001, S. 98–116

[Bishop 2006] Bishop, Christopher M.: *Pattern recognition and machine learning.* New York: Springer, 2006

[Blei u. a. 2004] Blei, David M.; Griffiths, Thomas L.; Jordan, Michael I. ; Tenenbaum, Joshua B.: Hierarchical topic models and the nested Chinese restaurant process. In: *Advances in Neural Information Processing Systems* Bd. 16, MIT Press, 2004 (NIPS '04), S. 106–114

[Blei u. Lafferty 2006] Blei, David M.; Lafferty, John D.: Dynamic Topic Models. In: *Proceedings of the 23rd International Conference on Machine Learning*, ACM, 2006 (ICML '06), S. 113–120

[Blei u. Lafferty 2007] Blei, David M.; Lafferty, John D.: A correlated topic model of Science. In: *The Annals of Applied Statistics* 1 (2007), Nr. 1, S. 17–35

[Blei u. a. 2003] Blei, David M.; Ng, Andrew Y. ; Jordan, Michael I.: Latent dirichlet allocation. In: *Journal of Machine Learning Research* 3 (2003), S. 993–1022

[Bonfadelli 2002] Bonfadelli, Heinz: *Medieninhaltsforschung: Grundlagen, Methoden, Anwendungen.* Konstanz: UVK Verlagsgesellschaft, 2002

[Bordag 2007] Bordag, Stefan: *Elements of knowledge-free and unsupervised lexical acquisition.* Leipzig, Universität Leipzig, Diss., 2007

[Bordag 2008] Bordag, Stefan: A comparison of co-occurrence and similarity measures as simulations of context. (2008), S. 52–63

[Bostock u. a. 2011] Bostock, Michael; Ogievetsky, Vadim ; Heer, Jeffrey: D^3 data-driven documents. In: *Visualization and Computer Graphics, IEEE Transactions on* 17 (2011), Nr. 12, S. 2301–2309

[Brandes 2001] Brandes, Ulrik: Drawing on Physical Analogies. 2025 (2001), S. 71–86

[Brinker 1988] Brinker, Klaus: *Linguistische Textanalyse: eine Einführung in Grundbegriffe und Methoden.* Berlin: Erich Schmidt Verlag, 1988

[Burghardt u. Wolff 2009] Burghardt, Manuel; Wolff, Christian: Stand off-Annotation für Textdokumente: Vom Konzept zur Implementierung (zur Standardisierung?). (2009), S. 53–59

[Busse 2009] Busse, Dietrich: *Semantik*. Stuttgart: UTB, 2009

[Ferrer i Cancho u. Solé 2001] Cancho, Ramon Ferrer i.; Solé, Richard V.: The small world of human language. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268 (2001), Nr. 1482, S. 2261–2265

[Danes u. Viehweger 1976] Danes, Frantisek; Viehweger, Dieter: *Probleme der Textgrammatik I.* Berlin: Akademie-Verlag, 1976

[Deerwester 1988] Deerwester, Scott: Improving Information Retrieval with Latent Semantic Indexing. In: *Proceedings of the 51st Annual Meeting of the American Society for Information Science*, 1988, S. 36–40

[Deerwester u. a. 1990] Deerwester, Scott; Dumais, Susan T.; Furnas, George W.; Landauer, Thomas K. ; Harshman, Richard: Indexing by latent semantic analysis. In: *Journal of the American Society for Information Science* 41 (1990), Nr. 6, S. 391–407

[Di Battista 1999] Di Battista, Giuseppe (Hrsg.): *Graph drawing: algorithms for the visualization of graphs.* Upper Saddle River: Prentice Hall, 1999

[van Dijk 1980] Dijk, Teun A.: *Textwissenschaft.* M. Niemeyer, 1980

[Dudenredaktion (Bibliographisches Institut) 2004] Dudenredaktion (Bibliographisches Institut): *Duden: die deutsche Rechtschreibung; auf der Grundlage der neuen amtlichen Rechtschreibregeln.* Berlin: Dudenverlag, 2004

[Dunning 1993] Dunning, Ted: Accurate Methods for the Statistics of Surprise and Coincidence. In: *Computational linguistics* 19 (1993), Nr. 1, S. 61–74

[DWDS] Berlin-Brandenburgische Akademie der Wissenschaften (Hrsg.): *DWDS / Suchergebnisse - Thema.* `http://www.dwds.de/?qu=Thema`. – Zugriff: 2014-12-20

[Eroms 1986] Eroms, Hans-Werner: *Funktionale Satzperspektive.* M. Niemeyer, 1986

[Evans 2014] Evans, Michael S.: A Computational Approach to Qualitative Analysis in Large Textual Datasets. In: *PLoS ONE* 9 (2014), Nr. 2

[Faruqui u. Padó 2010] Faruqui, Manaal; Padó, Sebastian: Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In: *Proceedings of the 10. Konferenz zur Verarbeitung Natürlicher Sprache.* Saarbrücken, Germany: GSCL, 2010 (KONVENS '10)

[Fellbaum 1998] Fellbaum, Christiane (Hrsg.): *WordNet: an electronic lexical database.* Cambridge, USA: MIT Press, 1998

[Fillmore 1968] Fillmore, Charles: The Case for Case. In: *Universals in Linguistic Theory*. Holt, Rinehart, and Winston, 1968

[Finkel u. a. 2005] Finkel, Jenny R.; Grenager, Trond ; Manning, Christopher: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL, 2005 (ACL '05), S. 363–370

[Firbas 1992] Firbas, Jan: *Functional sentence perspective in written and spoken communication*. Camebridge, UK: Cambridge University Press, 1992

[Fischer 2001] Fischer, Marc: *Produktlebenszyklus und Wettbewerbsdynamik: Grundlagen für die ökonomische Bewertung von Markteintrittsstrategien*. Wiesbaden: Deutscher Universitäts-Verlag, 2001

[Fleischer u. Hirsch 2001] Fleischer, Rudolf; Hirsch, Colin: Graph Drawing and Its Applications. (2001), S. 1–22

[Foulds u. a. 2013] Foulds, James; Boyles, Levi; DuBois, Christopher; Smyth, Padhraic ; Welling, Max: Stochastic Collapsed Variational Bayesian Inference for Latent Dirichlet Allocation. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2013 (KDD '13), S. 446–454

[Frege 1993] Frege, Gottlob: *Logische Untersuchungen* . Göttingen: Vandenhoeck und Ruprecht, 1993

[Fritz 1982] Fritz, Gerd: *Kohärenz: Grundfragen der linguistischen Kommunikationsanalyse*. Tübingen: Narr, 1982

[Früh 2001] Früh, Werner: Kategorienexploration bei der Inhaltsanalyse. Basiswissengeleitete offene Kategorienbildung (BoK). In: *Inhaltsanalyse: Perspektiven, Probleme, Potentiale*. Köln: Halem, 2001, S. 117–139

[Früh 2007] Früh, Werner: *Inhaltsanalyse : Theorie und Praxis*. Konstanz: UVK Verlagsgesellschaft, 2007

[Fukumoto u. Suzuki 2002] Fukumoto, Fumiyo; Suzuki, Yoshimi: Detecting Shifts in News Stories for Paragraph Extraction. In: *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, ACL, 2002 (COLING '02), S. 1–7

[Genzel 2005] Genzel, Dmitriy: A paragraph boundary detection system. In: *Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing*, Springer, 2005 (CICLing '05), S. 816–826

[Griffiths u. Steyvers 2004] Griffiths, Thomas L.; Steyvers, Mark: Finding scientific topics. In: *Proceedings of the National Academy of Sciences* Bd. 101, National Acad Sciences, 2004, S. 5228–5235

[Guardian] Guardian News and Media Limited (Hrsg.): *The Guardian Open Platform – Open Platform – The Guardian*. http://www.theguardian.com/open-platform. – Zugriff: 2014-12-20

[Handelsblatt] Handelsblatt GmbH (Hrsg.): *Monats-Chronik März 2011: Grünes Novum und Guttenberg-Aus*. http://www.handelsblatt.com/jahreswechsel/jahreswechsel-das-war-2011/monats-chronik-maerz-2011-gruenes-novum-und-guttenberg-aus/5931276.html. – Zugriff: 2014-12-20

[Hartung 2009] Hartung, Joachim: *Statistik : Lehr- und Handbuch der angewandten Statistik*. München: Oldenbourg, 2009

[Hastie u. a. 2001] Hastie, Trevor; Tibshirani, Robert ; Friedman, J. H.: *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. Berlin: Springer, 2001

[Hausser 2000] Hausser, Roland: *Grundlagen der Computerlinguistik: Mensch-Maschine-Kommunikation in natürlicher Sprache ; mit 772 Übungen*. Berlin: Springer, 2000

[Heinrich 2005] Heinrich, Gregor: Parameter estimation for text analysis / arbylon.net and Fraunhofer Computer Graphics Institute. Version: 2005. http://www.arbylon.net/publications/text-est.pdf. 2005. – Forschungsbericht. – Zugriff: 2014-12-20

[Heyer 2006] Heyer, Gerhard: *Text Mining: Wissensrohstoff Text : Konzepte, Algorithmen, Ergebnisse*. Dortmund: W3L-Verlag, 2006

[Heyer u. a. 2014] Heyer, Gerhard; Niekler, Andreas ; Wiedemann, Gregor: *Brauchen die Digital Humanities eine eigene Methodologie? Überlegungen zur systematischen Nutzung von Text Mining Verfahren in einem politikwissenschaftlichen Projekt*. Proceedings, 2014

[Hoffman u. a. 2013] Hoffman, M.; Blei, D.; Wang, Chong ; Paisley, John: Stochastic Variational Inference. In: *Journal of Machine Learning Research* 14 (2013), S. 1303–1347

[Hoffman u. a. 2010] Hoffman, Matthew; Bach, Francis R. ; Blei, David M.: Online learning for latent dirichlet allocation. In: *advances in neural information processing systems*, Curran Associates, Inc., 2010 (NIPS '10), S. 856–864

[Hofmann 1998] Hofmann, Thomas: Unsupervised Learning from Dyadic Data. In: *Computational Linguistics* 1198 (1998), Nr. 510, S. 466–472

[Hofmann 1999] Hofmann, Thomas: Probabilistic Latent Semantic Indexing. In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1999 (SIGIR '99), S. 50–57

[Hofmann u. a. 1999] Hofmann, Thomas; Puzicha, Jan ; Jordan, Michael I.: Learning from dyadic data. In: *Processing* pages (1999), S. 466–472

[Höft 1992] Höft, Uwe: *Lebenszykluskonzepte: Grundlage für das strategische Marketing- und Technologiemanagement.* Berlin: Erich Schmidt Verlag, 1992

[Hull 1996] Hull, David A.: Stemming Algorithms - A Case Study for Detailed Evaluation. In: *Journal of the American Society for Information Science* 47 (1996), S. 70–84

[Jackendoff 1990] Jackendoff, Ray: *Semantic structures.* MIT Press, 1990

[Keim 2010] Keim, Daniel: *Mastering the information age: solving problems with visual analytics.* Genf: Eurographics Association, 2010

[Kepplinger 2011] Kepplinger, HansMathias: Der Nachrichtenwert der Nachrichtenfaktoren. In: *Journalismus als Beruf.* Berlin: VS Verlag für Sozialwissenschaften, 2011, S. 61–75

[Kim u. Oh 2011] Kim, Dongwoo; Oh, Alice: Topic Chains for Understanding a News Corpus. (2011), S. 163–176

[Kolb 2005] Kolb, Steffen: *Mediale Thematisierung in Zyklen: Theoretischer Entwurf und empirische Anwendung.* Köln: Halem, 2005

[Koltsov u. a. 2014] Koltsov, Sergei; Koltsova, Olessia ; Nikolenko, Sergey I.: Latent dirichlet allocation: stability and applications to studies of user-generated content. In: *Proceedings of the Web Science Conference*, ACM, 2014 (WebSci '14), S. 161–165

[Krippendorff 2004] Krippendorff, Klaus: *Content analysis : an introduction to its methodology.* Thousand Oaks: Sage Publications, 2004

[Kumaran u. Allan 2005] Kumaran, Giridhar; Allan, James: Using names and topics for new event detection. In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ACL, 2005 (HLT '05), S. 121–128

[Kunze u. Lemnitzer 2002] Kunze, Claudia; Lemnitzer, Lothar: GermaNet - representation, visualization, application. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*, European Language Resources Association, 2002 (LREC '02)

[Lemke u. a. 2015] Lemke, Matthias; Niekler, Andreas; Schaal, Gary S. ; Wiedemann, Gregor: Content Analysis between Quality and Quantity. Fulfilling Blended-Reading Requirements for the Social Sciences with a Scalable Text Mining Infrastructure. In: *Datenbank Spektrum* 15 (2015), Nr. 1, S. 7–14

[Lemke u. Stulpe 2015] Lemke, Matthias; Stulpe, Alexander: Text und soziale Wirklichkeit. In: *Zeitschrift für germanistische Linguistik* 43 (2015), Nr. 1

[Lohmann u. a. 2009] Lohmann, Steffen; Ziegler, Jürgen ; Tetzlaff, Lena: Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration. In: *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I*, Springer, 2009 (INTERACT '09), S. 392–404

[Lötscher 1987] Lötscher, Andreas: *Text und Thema: Studien zur thematischen Konstituenz von Texten.* Tübingen: M. Niemeyer, 1987

[Lovins 1968] Lovins, Julie B.: *Development of a stemming algorithm.* Cambridge, USA: MIT Information Processing Group, Electronic Systems Laboratory, 1968

[Luhmann 1979] Luhmann, Niklas: öffentliche Meinung. (1979), S. 29–61

[Mackeldey 1987] Mackeldey, Roger: *Alltagssprachliche Dialoge: Kommunikative Funktionen und syntaktische Strukturen.* Leipzig: Verl. Enzyklopädie, 1987

[Manning 1999] Manning, Christopher: *Foundations of statistical natural language processing.* Cambridge, USA: MIT Press, 1999

[Manning u. a. 2008] Manning, Christopher D.; Raghavan, Prabhakar ; Schütze, Hinrich: *Introduction to information retrieval.* Cambridge, UK: Cambridge University Press, 2008

[Marchionini 2006] Marchionini, Gary: Exploratory search: from finding to understanding. In: *Communications of the ACM* 49 (2006), Nr. 4, S. 41–46

[Masada u. a. 2009] Masada, Tomonari; Takasu, Atsuhiro; Hamada, Tsuyoshi; Shibata, Yuichiro ; Oguri, Kiyoshi: Bag of Timestamps: A Simple and Efficient Bayesian Chronological Mining. (2009), S. 556–561

[McDowall 1992] McDowall, David (Hrsg.): *Interrupted time series analysis.* Thousand Oaks: Sage Publications, 1992

[Merten 1995] Merten, Klaus: *Inhaltsanalyse: Einführung in Theorie, Methode und Praxis.* Opladen: Westdt. Verl., 1995

[Merten 2001] Merten, Klaus: Konsensanalyse. Ein neues Instrument der Inhaltsanalyse. In: *Inhaltsanalyse: Perspektiven, Probleme, Potentiale.* Köln: Halem, 2001, S. 234–243

[Mikheev 1998] Mikheev, Andrei: Feature lattices for maximum entropy modelling. In: *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, ACL, 1998 (COLING '98), S. 848–854

[Miltner u. Waldherr 2013] Miltner, Peter; Waldherr, Annie: Themenzyklus der Kriegsberichterstattung – ein Phasenmodell. In: *Publizistik* 58 (2013), Nr. 3, S. 267–287

[Moretti 2005] Moretti, Franco: *Graphs, maps, trees: abstract models for a literary history.* New York: Verso, 2005

[Neuendorf 2002] Neuendorf, Kimberly A.: *The content analysis guidebook.* Thousand Oaks: Sage Publications, 2002

[Newman u. a. 2009] Newman, David; Asuncion, Arthur; Smyth, Padhraic ; Welling, Max: Distributed algorithms for topic models. In: *The Journal of Machine Learning Research* 10 (2009), S. 1801–1828

[Newman u. a. 2007] Newman, David; Smyth, Padhraic; Welling, Max ; Asuncion, Arthur U.: Distributed inference for latent dirichlet allocation. In: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2007 (NIPS '07), S. 1081–1088

[Niekler u. Jähnichen 2012] Niekler, Andreas; Jähnichen, Patrick: Matching Results of Latent Dirichlet Allocation for Text. In: *Proceedings of the 11th International Conference on Cognitive Modeling*, Universitätsverlag der TU Berlin, 2012 (ICCM '12), S. 317–322

[Niekler u. a. 2012] Niekler, Andreas; Jähnichen, Patrick ; Heyer, Gerhard: ASV Monitor: Creating Comparability of Machine Learning Methods for Content Analysis. In: *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2012*, Springer, 2012 (ECML-PKDD '12)

[Niekler u. a. 2014a] Niekler, Andreas; Wiedemann, Gregor; Dumm, Sebastian ; Heyer, Gerhard: *Creating dictionaries for argument identification by reference data.* Poster, 2014

[Niekler u. a. 2014b] Niekler, Andreas; Wiedemann, Gregor ; Heyer, Gerhard: Leipzig Corpus Miner - A Text Mining Infrastructure for Qualitative Data Analysis. In: *Terminology and Knowledge Engineering 2014*, 2014 (TKE '14)

[openNLP] The Apache Software Foundation (Hrsg.): *Apache OpenNLP - Welcome to Apache OpenNLP.* `http://opennlp.apache.org/`. – Zugriff: 2014-12-20

[Porteous u. a. 2008] Porteous, Ian; Newman, David; Ihler, Alexander; Asuncion, Arthur; Smyth, Padhraic ; Welling, Max: Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2008 (KDD '08). – ISBN 978–1–60558–193–4, S. 569–577

[Porter 1997] Porter, Martin F.: An algorithm for suffix stripping. In: *Readings in information retrieval*, Morgan Kaufmann Publishers Inc., 1997, S. 313–316

[Porter 2001] Porter, Martin F.: *Snowball: A language for stemming algorithms.* 2001

[Rauchenzauner 2008] Rauchenzauner, Elisabeth: *Schlüsselereignisse in der Medienberichterstattung.* Berlin: VS Verlag für Sozialwissenschaften, 2008

[Reynar 1998] Reynar, Jeffrey C.: *Topic segmentation: Algorithms and applications.* Philadelphia, University of Pennsylvania, Diss., 1998

[Roberts 1997] Roberts, Carl W.: *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts.* New York: Routledge, 1997

[Rosen-Zvi u. a. 2010] Rosen-Zvi, Michal; Chemudugunta, Chaitanya; Griffiths, Thomas; Smyth, Padhraic ; Steyvers, Mark: Learning author-topic models from text corpora. In: *ACM Transactions on Information Systems* 28 (2010), Nr. 1, S. 1–38

[Rössler 2005] Rössler, Patrick: *Inhaltsanalyse*. Konstanz: UVK Verlagsgesellschaft, 2005

[Sato u. a. 2012] Sato, Issei; Kurihara, Kenichi ; Nakagawa, Hiroshi: Practical collapsed variational bayes inference for hierarchical dirichlet process. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012 (SIGKDD '12), S. 105–113

[Saussure 2001] Saussure, Ferdinand d.: *Grundfragen der allgemeinen Sprachwissenschaft*. Berlin: W. de Gruyter, 2001

[Scharkow 2012] Scharkow, Michael: *Automatische Inhaltsanalyse und maschinelles Lernen* . Berlin, Univ. der Künste Berlin, Diss., 2012

[Schenk 2007] Schenk, Michael: *Medienwirkungsforschung*. Tübingen: Mohr Siebeck, 2007

[Schüßler 2008] Schüssler, Hans W. (Hrsg.): *Digitale Signalverarbeitung 1: Analyse diskreter Signale und Systeme*. Berlin: Springer, 2008

[Settles 2012] Settles, Burr: *Active Learning*. San Rafael: Morgan & Claypool Publishers, 2012

[Singhal u. a. 1996] Singhal, Amit; Buckley, Chris ; Mitra, Mandar: Pivoted Document Length Normalization. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1996 (SIGIR '96)

[Sommer 2014] Sommer, Katharina (Hrsg.): *Automatisierung in der Inhaltsanalyse* . Köln: Halem, 2014

[Spiegel] SPIEGEL ONLINE GmbH (Hrsg.): *Chronologie des Kriegs: Wie sich Libyen von Gaddafi befreite*. `http://www.spiegel.de/politik/ausland/chronologie-des-kriegs-wie-sich-libyen-von-gaddafi-befreite-a-792996.html`. – Zugriff: 2014-12-20

[Sporleder u. Lapata 2004] Sporleder, Caroline; Lapata, Mirella: Automatic Paragraph Identification: A Study across Languages and Domains. In: *Proceedings of EMNLP 2004*, Association for Computational Linguistics, 2004 (EMNLP '04), S. 72–79

[Steyvers u. Tenenbaum 2005] Steyvers, Mark; Tenenbaum, Joshua B.: The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. In: *Cognitive science* 29 (2005), Nr. 1, S. 41–78

[Stock 2007] Stock, Wolfgang G.: Themenentdeckung und -verfolgung und ihr Einsatz bei Informationsdiensten für Nachrichten. In: *Information - Wissenschaft & Praxis* 58 (2007), Nr. 1, S. 41–46

[Stone 1966] Stone, Philip J. (Hrsg.): *The general inquirer: A computer approach to content analysis*. Cambridge, USA: MIT Press, 1966

[Tagesspiegel 1] Verlag Der Tagesspiegel GmbH (Hrsg.): *Chronologie: Ein Jahr Revolution in Libyen.* `http://www.tagesspiegel.de/politik/chronologie-ein-jahr-revolution-in-libyen/6219342.html`. – Zugriff: 2014-12-20

[Tagesspiegel 2] Verlag Der Tagesspiegel GmbH (Hrsg.): *Fukushima. Eine Chronik der Ereignisse.* `http://www.tagesspiegel.de/politik/die-atomkatastrophe-in-japan-fukushima-eine-chronik-der-ereignisse/9038136.html`. – Zugriff: 2014-12-20

[Teh 2006] Teh, Yee W.: A hierarchical Bayesian language model based on Pitman-Yor processes. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL, 2006 (COLING ACL '06), S. 985–992

[Teh u. Jordan 2010] Teh, Yee W.; Jordan, Michael. I.: Hierarchical Bayesian Nonparametric Models with Applications. In: Hjort, N. (Hrsg.); Holmes, C. (Hrsg.); Müller, P. (Hrsg.) ; Walker, S. (Hrsg.): *Bayesian Nonparametrics: Principles and Practice.* Cambridge University Press, 2010

[Teh u. a. 2007] Teh, Yee W.; Kurihara, Kenichi ; Welling, Max: Collapsed variational inference for HDP. In: *Advances in neural information processing systems*, Curran Associates, Inc., 2007 (NIPS '07), S. 1481–1488

[Teh u. a. 2006] Teh, Yee W.; Newman, David ; Welling, Max: A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: *Advances in neural information processing systems*, Curran Associates, Inc., 2006 (NIPS '06), S. 1353–1360

[Thomas u. Cook 2005] Thomas, J. J.; Cook, K. A.: *Illuminating the path: the research and development agenda for visual analytics.* New York: IEEE Computer Society Press, 2005

[Top 2006] Top, Jasmin: *Konsensanalyse: ein neues Instrument der Inhaltsanalyse: theoretische Fundierung und empirische Kalibrierung.* Norderstedt: Books on Demand, 2006

[Tukey 1977] Tukey, John W.: *Exploratory data analysis.* Reading: Addison-Wesley Pub. Co, 1977 (Addison-Wesley series in behavioral science)

[uima] The Apache Software Foundation (Hrsg.): *Apache UIMA - Welcome to the Apache UIMA project.* `https://uima.apache.org/`. – Zugriff: 2014-12-20

[Wang u. a. 2011] Wang, Chong; Paisley, John W. ; Blei, David M.: Online Variational Inference for the Hierarchical Dirichlet Process. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, JMLR .org, 2011 (AISTATS '11), S. 752–760

[Wang u. a. 2007] Wang, Chong; Wang, Jinggang; Xie, Xing ; Ma, Wei-Ying: Mining Geographic Knowledge Using Location Aware Topic Model. In: *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, ACM, 2007 (GIR '07), S. 65–70

[Wang u. McCallum 2006] Wang, Xuerui; McCallum, Andrew: Topics over Time: A non-Markov Continuous-time Model of Topical Trends. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2006 (KDD '06), S. 424–433

[Wang u. a. 2009] Wang, Yi; Bai, Hongjie; Stanton, Matt; Chen, Wen-Yen ; Chang, Edward Y.: PLDA: Parallel Latent Dirichlet Allocation for Large-Scale Applications. In: *Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management*, Springer-Verlag, 2009 (AAIM '09), S. 301–314

[Watts u. Strogatz 1998] Watts, Duncan J.; Strogatz, Steven H.: Collective dynamics of small-world networks. In: *Nature* 393 (1998), Nr. 6684, S. 440–442

[West 2001] West, Mark D. (Hrsg.): *Applications of computer content analysis* . Westport: Ablex Pub., 2001

[White u. Roth 2009] White, Ryen W.; Roth, Resa A.: Exploratory search beyond the query-response paradigm. In: *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1 (2009), Nr. 1, S. 1–98

[Wiedemann u. a. 2013] Wiedemann, Gregor; Lemke, Matthias ; Niekler, Andreas: Postdemokratie und Neoliberalismus. Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949-2011. In: *Zeitschrift für Politische Theorie (ZPTh)* 4 (2013), Nr. 1, S. 99–115

[Wiedemann u. Niekler 2014] Wiedemann, Gregor; Niekler, Andreas: Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries. In: *Terminology and Knowledge Engineering 2014*, 2014 (TKE '14)

[Wiki 1] Wikimedia Foundation Inc. (Hrsg.): *Chronik der Nuklearkatastrophe von Fukushima*. `http://de.wikipedia.org/wiki/Chronik_der_Nuklearkatastrophe_von_Fukushima`. – Zugriff: 2014-12-20

[Wiki 2] Wikimedia Foundation Inc. (Hrsg.): *Chronik des Bürgerkriegs in Libyen*. `http://de.wikipedia.org/wiki/Chronik_des_B{ü}rgerkriegs_in_Libyen`. – Zugriff: 2014-12-20

[Wiki 3] Wikimedia Foundation Inc. (Hrsg.): *Chronologie der Katastrophe in Japan von 2011*. `http://de.wikipedia.org/wiki/Chronologie_der_Katastrophe_in_Japan_von_2011`. – Zugriff: 2014-12-20

[Wiki 4] Wikimedia Foundation Inc. (Hrsg.): *Thema – Wikipedia*. `http://de.wikipedia.org/wiki/Thema`. – Zugriff: 2014-12-20

[Wirth u. Lauf 2001] Wirth, Werner (Hrsg.); Lauf, Edmund (Hrsg.): *Inhaltsanalyse: Perspektiven, Probleme, Potentiale*. Köln: Halem, 2001

[wortschatz 1] Abteilung Automatische Sprachverarbeitung – Universität Leipzig (Hrsg.): *Wortschatz*. `http://wortschatz.uni-leipzig.de/`. – Zugriff: 2014-12-20

[wortschatz 2] Abteilung Automatische Sprachverarbeitung – Universität Leipzig (Hrsg.): *Wortschatz - Result - Thema*. `http://wortschatz.uni-leipzig.de/cgi-portal/de/wort_www?site=10&Wort_id=21731&Wort=Thema&stpw=0&verweise=7&kanz=32`. – Zugriff: 2014-12-20

[Yan u. a. 2009] Yan, Feng; Xu, Ningyi ; Qi, Yuan: Parallel inference for latent dirichlet allocation on graphics processing units. In: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2009 (NIPS '09), S. 2134–2142

[Zeit] ZEIT ONLINE GmbH (Hrsg.): *Chronologie: Vom Aufstand zur Intervention*. `http://www.zeit.de/news-032011/20/iptc-bdt-20110319-450-29352636xml`. – Zugriff: 2014-12-20

Curriculum Vitae

Andreas Niekler

Josephstraße 33
04177 Leipzig, Deutschland

Email: andreas.niekler@gmx.de

geb: Juli 21, 1979 – Torgau, Deutschland
Nationalität: deutsch

Derzeitige Tätigkeit

wissenschaftlicher Mitarbeiter, Abteilung Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig, Augustusplatz 10, 04109 Leipzig, Germany

Arbeitsgebiete

Text Mining, Maschinelles Lernen, Inhaltsanalyse, Informationssysteme, Visualisierung, Kommunikationsforschung.

Ausbildung

Abitur, Berufliches Schulzentrum 6, Leipzig	2000-2001
Dipl. Ing. in Medientechnik, HTWK Leipzig, Deutschland	2001-2007
BSc in Media Technology, University of West Scotland, Paisley, Schottland	2003-2004

beruflicher Werdegang

1996-1999 Ausbildung zum Vermessungstechniker

2005-2015 Freiberuf Programmierung, Beratung

2007-2009 Laboringenieur, HTWK Leipzig

2009-2012 wissenschaftlicher Mitarbeiter, HTWK Leipzig

2012-2015 wissenschaftlicher Mitarbeiter, Universität Leipzig

Berufserfahrung, Projekte, Kooperationen

2005-2009 Betreuung, Programmierung, Kozeptionierung einer Plattform für Hostel- und Hotelbuchung. (GOMIO.COM Barcelona)

2005-2009 Betreuung, Programmierung, Kozeptionierung von CMS Systemen (Circus Hostel Berlin, Lubok Verlag)

2007-2010 Umsetzung eines IPTV Systems für das Campus-TV der HTWK Leipzig

2009-2010 Umsetzung einer Streaming-Plattform für das Campus-TV der HTWK Leipzig

2012-2014 Umsetzung und Betreuung eines Verwaltungssystems für textbasierte Umfragedaten (Siemens AG Global Shared Services Mannheim)

2013-2014 Beratung und Umsetzung einer Analyse für den Vergleich des Begriffs “Diabetes” in deutschen und chinesischen Internet-Quellen (Li Xiguang Tsinghua Universität Beijing)

2014-2015 Beratung und Umsetzung einer Analysesoftware für die Messung von ausgedrückten Emotionen in Textdaten (m-result Mainz)

2014-2015 Forschungsantrag und Projektbegleitung für ein kooperatives Forschungsprojekt im Rahmen der Arbeitsgemeinschaft industrieller Forschungsvereinigungen (AiF) (Fach-

hochschule Hof, puls Marketing GmbH Nürnberg)

Umsetzung einer Architektur für die Inhaltsanalyse großer Textmengen im Verbundprojekt *Postdemokratie und Neoliberalismus: Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949-2011* (Universität Leipzig) 2012-2015

Lehrtätigkeit

Medienneutrales Publizieren, XML (Leipzig School of Media)	2007-2015
Seminar Linguistische Informatik (Universität Leipzig)	2013-2015
Seminar Text Mining (Universität Leipzig)	2013-2015

Veröffentlichungen, Vorträge

2015

- Wiedemann, Gregor; Niekler, Andreas: *Analyse qualitativer Daten mit dem „Leipzig Corpus Miner“* In: *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*, Reihe: Kritische Studien zur Demokratie, VS Verlag für Sozialwissenschaften, 2015.

- Dumm, Sebastian; Niekler, Andreas: *Methoden, Qualitätssicherung und Forschungsdesign. Diskurs- und Inhaltsanalyse zwischen Sozialwissenschaften und automatischer Sprachverarbeitung* In: *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*, Reihe: Kritische Studien zur Demokratie, VS Verlag für Sozialwissenschaften, 2015.

- Matthias Lemke, Andreas Niekler, Gary S. Schaal, Gregor Wiedemann: *Content Analysis between Quality and Quantity. Fulfilling Blended-Reading Requirements for the Social Sciences with a Scalable Text Mining Infrastructure*, Datenbank Spektrum, 2015

2014

- Andreas Niekler, Gregor Wiedemann, Sebastian Dumm, Gerhard Heyer: *Creating dictionaries for argument identification by reference data.* In: *1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014)*, 2014

- Gerhard Heyer, Andreas Niekler, Gregor Wiedemann: *Brauchen die Digital Humanities eine eigene Methodologie? Überlegungen zur systematischen Nutzung von Text Mining Verfahren in einem politikwissenschaftlichen Projekt.* In: *1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014)*, 2014

- Hänig, C., Niekler, A. und Wünsch, C.: *PACE Corpus: a Multilingual Corpus of Polarity-Annotated Textual Data from the Domains Automotive and Cellphone.* In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), 2014

- Daniel Isemann, Andreas Niekler, Benedict Preßler, Frank Viereck und Gerhard Heyer: *OCR of Legacy Documents as a Building Block in Industrial Disaster Prevention.* In: *DIMPLE Workshop on Disaster Management and Principled Large-scale information Extraction for and post emergency logistics*, LREC 2014, 2014

- Andreas Niekler, Gregor Wiedemann und Gerhard Heyer: *Leipzig Corpus Miner - A Text Mining Infrastructure for Qualitative Data Analysis.* In: *Terminology and Knowledge Engineering 2014 (TKE 2014)*, Berlin, 2014

- Gregor Wiedemann und Andreas Niekler: *Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries.* In: *Terminology and Knowledge Engineering 2014 (TKE 2014)*, Berlin, 2014

- Gregor Wiedemann und Andreas Niekler: *Introductory talk for the 3 day workshop* Vortrag: *Text Mining in der Politikwissenschaft*, Helmut-Schmidt-University, Hamburg, 2014

- Gregor Wiedemann und Andreas Niekler: *Series of lectures and tutorials for the 5 day workshop* Vortrag: *Basismodul II – Text Mining for Social Scientists*, Gesis – Leibniz Institut für Sozialwissenschaften, Köln, 2014

- Andreas Niekler: *Leipzig Corpus Miner - A Text Mining Infrastructure for Qualitative Data Analysis* Vortrag: *Terminology and Knowledge Engineering 2014 conference (TKE '14)*, Berlin, 2014

- Gregor Wiedemann und Andreas Niekler: *Tracking down economization. Large scale text analysis for political sciences* Vortrag: *Workshop mit Gregory Crane / Perseus project*, Tufts University, Boston, 2014

- Gregor Wiedemann und Andreas Niekler: *Tracking down economization. Large scale text analysis for political sciences* Vortrag: *Liechtenstein Institute on Self-Determination (LISD)*, Princeton University, Princeton, 2014

2013

- Gregor Wiedemann, Matthias Lemke und Andreas Niekler: *Postdemokratie und Neoliberalismus. Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949-2011.* In: *Zeitschrift für politische Theorie*, 2013

- Andreas Niekler: *Exemplarische Studie zur journalistischen Sicht auf Europa* Vortrag: *Tagung Europa und der Journalismus - Zwischen Stimmungsmache und Aufklärung*, Hamburg Media School, 2013

- Andreas Niekler: *ASVMonitor und exemplarische Analysen zum Thema "Öffentliches Vertrauen"* Vortrag: *Wandel und Messbarkeit des öffentlichen Vertrauens im Zeitalter des Web 2.0*, Leipzig School of Media, 2013

- Gregor Wiedemann und Andreas Niekler: *Tracking down economization. Large scale text analysis for political sciences* Vortrag: *VisArgue Kickoff-Workshop*, Konstanz, 2013

- Gregor Wiedemann und Andreas Niekler: *Text Mining for Large Scale Corpus Analyses in the eHumanities* Vortrag: *Aktuelle Entwicklungen der sozialwissenschaftlichen quantitativen Diskursforschung/Textinhaltsanalyse im deutschsprachigen Raum*, Bremen, 2014

2012

- Andreas Niekler und Patrick Jähnichen: *Matching Results of Latent Dirichlet Allocation for Text.* In: *Proceedings of ICCM 2012, 11th International Conference on Cognitive Modeling*, Berlin, 2012

- C. Rohrdantz, A. Niekler, A. Hautli, M. Butt und D. A. Keim: *Lexical Semantics and Distribution of Suffixes - A Visual Analysis.* In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, 2012*

- Andreas Niekler, Patrick Jähnichen und Gerhard Heyer: *ASV Monitor: Creating Comparability of Machine Learning Methods for Content Analysis.* In: *Proceedings of the ECML-PKDD 2012*, Bristol, 2012

- Gregor Wiedemann und Andreas Niekler: *Der Ökonomisierung auf der Spur: Text Mining zur Analyse großer Zeitungskorpora* Vortrag: *Politikwissenschaft und die Methoden der eHumanities*, Helmut-Schmidt-University, Hamburg, 2014

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

_____

(Ort, Datum)

_____

(Unterschrift)